ΑΥΤΟΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΧΙΟΥΜΟΡΙΣΤΙΚΩΝ ΥΠΟΤΙΤΛΩΝ

Στυλιανός - Γεώργιος Μαμμάς

Πολυτεχνείο Κρήτης

Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

7 Ιουλίου 2010

Εγκρίθηκε από την τριμελή επιτροπή:

1. Αν. καθ. Αλέξανδρος Ποταμιάνος (Επιβλέπων)

2. Καθ. Βασίλειος Διγαλάκης

3. Αν. καθ. Ευριπίδης Πετράκης

Περιεχόμενα

Κατάλογος Σχημάτων

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Αλέξανδρο Ποταμιάνο, για την πολύτιμη καθοδήγηση του κατά την διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Επίσης ευχαριστώ θερμά τον μεταπτυχιακό φοιτητή κ. Νίκο Μαλανδράκη, για την αμέριστη βοήθεια του.

Εισαγωγή

Η έρευνα για την αυτόματη αναγνώριση και παραγωγή χιουμοριστικών εκφράσεων από υπολογιστή, βρίσκεται ακόμη σε πρώιμο στάδιο. Ο λόγος είναι ότι στην χιουμοριστική γλώσσα χρησιμοποιούνται πολύπλοκες, διφορούμενες, και ασυνεπείς συντακτικές και σημασιολογικές δομές, οι οποίες απαιτούν βαθιά σημασιολογική ερμηνεία - εξήγηση. Ένας σημαντικός λόγος μοντελοποίησης του χιούμορ είναι η εφαρμογή του στην διεπαφή χρήστη - υπολογιστή, κάνοντας έτσι την αλληλεπίδραση πιο φυσική και ενδιαφέρουσα για τον χρήστη.

Ο σκοπός της παρούσας εργασίας είναι η δημιουργία ενός συστήματος, το οποίο θα αναγνωρίζει αυτόματα τους χιουμοριστικούς υπότιτλους με βάση γλωσσικά χαρακτηριστικά. Οι υπότιτλοι που χρησιμοποιούνται σαν δεδομένα, προέρχονται από 23 επεισόδια της 4ης σεζόν της χιουμοριστικής σειράς Friends. Η εργασία αποτελείται από δύο μέρη.

Στόχος του πρώτου μέρους είναι η δημιουργία μίας συλλογής από χιουμοριστικούς και μη χιουμοριστικούς υπότιτλους. Γίνεται διαχωρισμός των χιουμοριστικών από τους μη χιουμοριστικούς υπότιτλους, για όλα τα επεισόδια που προαναφέραμε. Επισημειώνονται τα διαστήματα μηχανικού/μη μηχανικού γέλιου για τα τρία πρώτα επεισόδια, και στην συνέχεια χρησιμοποιώντας τεχνικές εκμάθησης με επίβλεψη και το μοντέλο μείξης κανονικών κατανομών (mixture of Gaussians), αναγνωρίζονται τα διαστήματα αυτά και στα υπόλοιπα (20) επεισόδια. Στην συνέχεια χρησιμοποιείται ένα απλό σχήμα επισήμανσης των χιουμοριστικών υποτίτλων, στο οποίο ένας υπότιτλος θεωρείται χιουμοριστικός, εάν βρίσκεται αμέσως πριν από ένα διάστημα μηχανικού γέλιου (artificial laugh). Έτσι με το πέρας αυτού του σκέλους της εργασίας, δημιουργείται ένας ικανός όγκος από χιουμοριστικά και μη χιουμοριστικά γλωσσικά δεδομένα (υπότιτλοι).

Στο δεύτερο μέρος της εργασίας, χρησιμοποιούνται τα γλωσσικά δεδομένα που δημιουργήθηκαν στο πρώτο μέρος, ώστε να γίνει η αναγνώριση των χιουμοριστικών υποτίτλων με βάση γλωσσικά χαρακτηριστικά. Για τον σκοπό αυτό χρησιμοποιείται το N-gram πιθανοτικό μοντέλο, το μοντέλο του διανυσματικού χώρου και η λανθάνουσα σημασιολογική ανάλυση. Επίσης γίνεται μελέτη της ορθότητας των ταξινομητών, αναφορικά με την απόρριψη υποτίτλων για τους οποίους υπάρχει υψηλή αβεβαιότητα για την κατηγορία στην οποία ανήκουν. Τέλος γίνεται επιλογή χαρακτηριστικών με βάση την αμοιβαία πληροφορία και μελετάται η ορθότητα των ταξινομητών, αναφορικά με το πλήθος των χαρακτηριστικών που χρησιμοποιούνται στην εκπαίδευση των μοντέλων.

Διάρθρωση της εργασίας

Τα τέσσερα πρώτα κεφάλαια προσφέρουν, στον αναγνώστη, το απαραίτητο θεωρητικό υπόβαθρο για την κατανόηση της εργασίας. Στο κεφάλαιο 5 παρουσιάζεται σχετική έρευνα με την εργασία μας. Στα κεφάλαια 6, 7 παρουσιάζεται η εργασία μας. Το κεφάλαιο 6 είναι το πρώτο σκέλος της εργασίας και αφορά την δημιουργία των γλωσσικών δεδομένων. Το κεφάλαιο 7 είναι το δεύτερο σκέλος της εργασίας και παρουσιάζει την αυτόματη αναγνώριση των χιουμοριστικών υποτίτλων με βάση γλωσσικά χαρακτηριστικά. Τέλος στο κεφάλαιο 8 παρουσιάζονται τα συμπεράσματα που βγαίνουν από την εργασία και ο μελλοντικός σχεδιασμός για περαιτέρω βελτίωση των αποτελεσμάτων.

Μέρος Ι

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Κεφάλαιο 1

Το μοντέλο μείξης κανονικών κατανομών

Το μοντέλο της μείξης κανονικών κατανομών (Gaussian mixture model) είναι ένα πιθανοτικό μοντέλο στο οποίο η κατανομή της πιθανότητας περιγράφεται από ένα γραμμικό συνδυασμό κανονικών κατανομών. Έστω ένα σύνολο δεδομένων $D = \{\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N)\}$, όπου $\mathbf{x}(i)$ είναι τα d-διάστατα διανύσματα των μετρήσεων. Υποθέτουμε ότι τα δεδομένα παράγονται από μία υποκείμενη κατανομή $p(\mathbf{x})$, και ότι είναι αμοιβαίως ανεξάρτητα μεταξύ τους (Independent Identically Distribution - IID). Επιπλέον υποθέτουμε ότι υπάρχουν K κανονικές κατανομές στο μοντέλο μείξης κανονικών κατανομών. Σύμφωνα με τα παραπάνω μπορούμε να γράψουμε ότι:

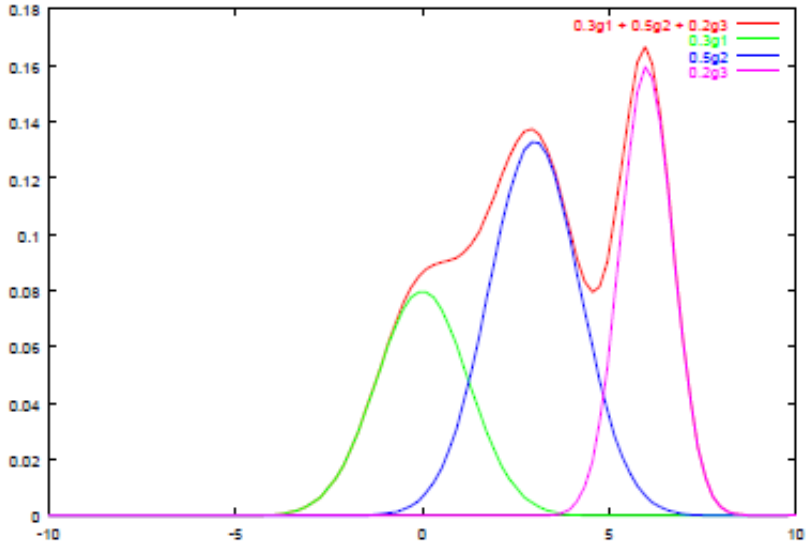
$$p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \alpha_k \, p_k(\mathbf{x}|\theta_k) \tag{1.1}$$

Στον παραπάνω τύπο τα α_k είναι τα βάρη των κατανομών, και ισχύει ότι:

$$\sum_{k=1}^{K} \alpha_k = 1 \tag{1.2}$$

Το Θ συμβολίζει τις παραμέτρους του μοντέλου:

$$\Theta = \{\alpha_1, ..., \alpha_K, \theta_1, ..., \theta_K\} \tag{1.3}$$



Σχήμα 1.1: Παράδειγμα μείξης τριών κανονικών κατανομών

Εφόσον κάθε συνιστώσα (component) της μείξης είναι μία πολυδιάστατη κανονική κατανομή, με τις δικές της παραμέτρους $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, έχουμε ότι:

$$p_k(\mathbf{x}|\theta_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)} \tag{1.4}$$

Τέλος η πιθανότητα που έχει ένα δεδομένο (data point) $\mathbf{x}(i)$, να ανήκει στην κατανομή k, δίνεται από τον ακόλουθο τύπο:

$$w_{ik} = p(C = k|\mathbf{x}(i), \Theta) = \frac{p_k(\mathbf{x}(i)|\theta_k).\alpha_k}{\sum_{m=1}^{K} p_m(\mathbf{x}(i)|\theta_m).\alpha_m}, \; 1 \leq k \leq K, \; 1 \leq i \leq N \tag{1.5}$$

1.1 Ο αλγόριθμος Expectation-Maximization

Σε αυτήν την ενότητα θα περιγράψουμε τον EM (Expectation - Maximization) αλγόριθμο για την περίπτωση της μείξης κανονικών κατανομών. Ο αλγόριθμος EM περιλαμβάνει δύο βήματα:

E-step: Δηλώνουμε τις τρέχουσες τιμές των παραμέτρων μας ως Θ. Με βάση την εξίσωση 1.5 υπολογίζουμε τις τιμές των w_{ik}, για όλα τα δεδομένα $(\mathbf{x}(i),\ 1 \leq i \leq N)$ και για όλες τις συνιστώσες της μείξης κανονικών κατανομών $(1 \leq k \leq K)$. Να σημειωθεί ότι από τον τον τρόπο που ορίστηκαν τα w_{ik} (τύπος 1.5) ισχύει ότι $\sum_{k=1}^{K} w_{ik} = 1$. Έτσι δημιουργείται ένας πίνακας N x K ο οποίος περιέχει τα w_{ik}. Όπως είναι φανερό κάθε γραμμή του πίνακα αυτού αθροίζει στο 1.

M-step: Στο βήμα αυτό χρησιμοποιούνται οι τιμές των w_{ik} που υπολογίστηκαν στο στο E-step, για να υπολογιστούν οι καινούργιες τιμές των παραμέτρων. Ειδικότερα για τα καινούργια βάρη των κατανομών έχουμε:

$$\alpha_k^{new} = \frac{1}{N} \sum_{i=1}^{N} w_{ik}\ ,\ 1 \leq k \leq K \tag{1.6}$$

Οι καινούργιες μέσες τιμές υπολογίζονται ως σταθμισμένοι μέσοι όροι των δεδομένων:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{\sum_{i=1}^{N} w_{ik}} \sum_{i=1}^{N} w_{ik}\,.\,\mathbf{x}(i),\ \ 1 \leq k \leq K \tag{1.7}$$

Θυμίζουμε ότι τα $\boldsymbol{\mu}_k^{new}$, $\mathbf{x}(i)$ είναι d-διάστατα διανύσματα. Όσον αφορά τους καινούργιους πίνακες συνδιασπορών ($\boldsymbol{\Sigma}_k^{new}$), υπολογίζονται ως εξής:

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{\sum_{i=1}^{N} w_{ik}} \sum_{i=1}^{N} w_{ik}\,.\,(\mathbf{x}(i) - \boldsymbol{\mu}_k^{new})(\mathbf{x}(i) - \boldsymbol{\mu}_k^{new})^T,\ \ 1 \leq k \leq K \tag{1.8}$$

Ο παραπάνω τύπος είναι παρόμοιος με τον κλασσικό τύπο που χρησιμοποιείται για τον υπολογισμό ενός πίνακα συνδιασπορών, με την διαφορά ότι η συνεισφορά κάθε

δεδομένου σταθμίζεται με ένα βάρος w_{ik}. Σημειώνουμε ότι κάθε πίνακας συνδιασπορών είναι διάστασης d x d.

Οι εξισώσεις του βήματος M-step πρέπει να υπολογίζονται με την σειρά που παρουσιάστηκαν: Πρώτα υπολογίζονται τα α_k, μετά τα K καινούργια $\boldsymbol{\mu}_k$, και τέλος τα K καινούργια $\boldsymbol{\Sigma}_k$.

Έχοντας υπολογίσει όλες τις καινούργιες παραμέτρους, το M-step έχει ολοκληρωθεί. Μαζί, τα βήματα E-step και M-step, αποτελούν μία επανάληψη. Μόλις τελειώσει μία επανάληψη, ξεκινάει μια καινούργια, επιστρέφοντας στο E-step και υπολογίζοντας τις καινούργιες τιμές των παραμέτρων. Ο αλγόριθμος σταματάει όταν ξεπεραστεί το όριο επαναλήψεων που έχουμε θέσει ή αν ικανοποιηθεί το κριτήριο σύγκλισης.

Αρχικοποίηση και κριτήριο σύγκλισης

Για να ξεκινήσει ο αλγόριθμος EM πρέπει να αρχικοποιηθούν οι τιμές των παραμέτρων και των βαρών που έχει κάθε συνιστώσα-κατανομή. Οι αρχικές παράμετροι ή τα αρχικά βάρη, μπορούν να επιλεγούν είτε τυχαία (π.χ επιλογή K τυχαίων δεδομένων σαν αρχικές μέσες τιμές και αρχικοποίηση των K πινάκων συνδιασπορών με τον πίνακα συνδιασπορών όλων των δεδομένων), είτε μέσω κάποιας μεθόδου, όπως τη χρησιμοποίηση του αλγόριθμου K-means, για τον διαχωρισμό των δεδομένων σε ομάδες (clusters) και τον προσδιορισμό των βαρών με βάση τα μέλη κάθε ομάδας.

Η σύγκλιση του αλγορίθμου EM ανιχνεύεται ως εξής: Υπολογίζεται ο λογάριθμος της πιθανότητας (log-likelihood) των δεδομένων μετά από κάθε επανάληψη. Αποδεικνύεται ότι η πιθανότητα των δεδομένων δεν μειώνεται από επανάληψη σε επανάληψη στον αλγόριθμο EM [18] . Στην ουσία επιτυγχάνεται ένα τοπικό μέγιστο της τιμής της πιθανότητας των δεδομένων. Όταν δεν υπάρχει σημαντική διαφορά στην στην τιμή του λογαρίθμου της πιθανότητας των δεδομένων (η οποία ορίζεται με κάποιο κατώφλι) ανάμεσα σε δύο επαναλήψεις, σημαίνει ότι το κριτήριο σύγκλισης έχει ικανοποιηθεί και ο αλγόριθμος τερματίζει. Σημειώνουμε ότι ο λογάριθμος της πιθανότητας των δεδομένων (υπό την IID θεώρηση), ορίζεται ως εξής:

$$\log l(\Theta) = \sum_{i=1}^{N} \log p(\mathbf{x}(i)|\Theta) \tag{1.9}$$

, όπου η $p(\mathbf{x}(i)|\Theta)$ είναι η εξίσωση του μοντέλου μείξης κανονικών κατανομών με τις K συνιστώσες που περιγράψαμε προηγουμένως.

Αναφορές

Ο αναγνώστης μπορεί να ανατρέξει στις πηγές [12], [16], [17], [18], για επιπλέον πληροφορίες σχετικά με το μοντέλο μείξης κανονικών κατανομών και τον αλγόριθμο EM.

Κεφάλαιο 2

N-gram πιθανοτικό μοντέλο

Σε αυτό το κεφάλαιο παρουσιάζονται έννοιες οι οποίες θα είναι χρήσιμες στον αναγνώστη για την κατανόηση της εργασίας μας. Στην ενότητα 2.1 εισάγουμε τον αναγνώστη στις συλλογές κειμένου, καθώς και σε κάποια θέματα σχετικά με τον χειρισμό των λέξεων αναλόγως με την εφαρμογή. Αργότερα, στην ενότητα 2.2, γίνεται μια εισαγωγή σε γλωσσικά πιθανοτικά μοντέλα. Τέλος, στις ενότητες 2.3 και 2.4 αντίστοιχα, εμβαθύνουμε σε αυτά τα μοντέλα με την παρουσίαση της εξομάλυνσης (smoothing) και του backoff. Η βασική πηγή πληροφοριών για το παρόν κεφάλαιο είναι το βιβλίο Speech and Language Processing, των D. Jurafsky, J. H. Martin. [10]

2.1 Εισαγωγή στις συλλογές κειμένου

Η στατιστική επιστήμη στην περιοχή της φυσικής γλώσσας βασίζεται σε συλλογές (corpora) κειμένου ή ομιλίας ανάλογα με την εφαρμογή. Για τον υπολογισμό της πιθανότητας που έχει κάθε λέξη, μετράμε τον αριθμό των εμφανίσεων κάθε λέξης στο training corpus. Για παράδειγμα παραθέτουμε την πρόταση (2.1):

They picnicked by the pool, then lay back on the grass and looked at the stars. (2.1)

Πόσες λέξεις περιέχει η παραπάνω πρόταση; Καταρχήν υπάρχουν δύο περιπτώσεις. Στην πρώτη περίπτωση συμπεριλαμβάνονται τα σημεία στίξης στο μέτρημα, αφού θεωρούνται σαν λέξεις (οπότε έχουμε 18 λέξεις). Στην δεύτερη περίπτωση

τα σημεία στίξης δεν λαμβάνονται υπόψιν (οπότε η πρόταση περιέχει 16 λέξεις). Αναλόγως με την εφαρμογή επιλέγεται αν θα αγνοηθούν ή όχι τα σημεία στίξης. Σε μερικές εφαρμογές όπως grammar-checking, spelling error detection, είναι απαραίτητο να συμπεριληφθούν τα σημεία στίξης. Η πιθανότητα μίας λέξης w είναι το κλάσμα του αριθμού των εμφανίσεων της λέξης προς τον συνολικό αριθμό των λέξεων:

$$p(w) = \frac{occurrences\ of\ word\ w}{number\ of\ words} \tag{2.2}$$

Έτσι για παράδειγμα η πιθανότητα της λέξης "grass" στην πρόταση 2.1, όταν λαμβάνονται υπόψιν τα σημεία στίξης, είναι $P(\text{“}grass\text{”}) = 1/18$. Όταν αγνοούνται είναι $P(\text{“}grass\text{”}) = 1/16$.

Ένα άλλο ερώτημα είναι το εξής: Είναι οι λέξεις με κεφαλαία γράμματα ίδιες με τις αντίστοιχες λέξεις γραμμένες με μικρά γράμματα (π.χ They και they); Στις περισσότερες εφαρμογές τέτοιες λέξεις θεωρούνται ίδιες. Υπάρχουν όμως και case-sensitive εφαρμογές (π.χ spelling error correction) στις οποίες τέτοιες λέξεις χειρίζονται με διαφορετικό τρόπο.

Πως πρέπει να χειριζόμαστε λέξεις οι οποίες βρίσκονται σε διαφορετική κλίση (π.χ work, works); Ξανά αυτό εξαρτάται από την εφαρμογή. Στα περισσότερα συστήματα τέτοιες λέξεις χειρίζονται σαν ξεχωριστές λέξεις. Αυτή η τακτική δεν είναι κατάλληλη σε πολλά συστήματα, στα οποία θα θέλαμε λέξεις όπως work, works, να αποτελούν στιγμιότυπα (instances) μιας αφηρημένης λέξης (lemma). Το λήμμα (lemma) είναι ένα σύνολο από λέξεις οι οποίες ανήκουν στο ίδιο μέρος του λόγου, έχουν ίδια ρίζα, και την ίδια λεκτική έννοια.

Κλείνοντας να σημειώσουμε ότι στην συνέχεια του κεφαλαίου χρησιμοποιούμε τον όρο types, για να δηλώσουμε το πλήθος των διαφορετικών λέξεων στο corpus. Το μέγεθος του λεξικού (vocabulary size) μίας εφαρμογής ισούται με το πλήθος των διαφορετικών λέξεων (word types) στο training corpus. Όπως είναι φανερό το μέγεθος του λεξικού είναι αρκετά μικρότερο από το συνολικό αριθμό των λέξεων (word tokens) του corpus, αφού πολλές λέξεις επαναλαμβάνονται. Για παράδειγμα η πρόταση 2.1 έχει 14 word types και 16 word tokens (δεν λαμβάνονται υπόψιν τα σημεία στίξης στο μέτρημα).

2.2 N-grams χωρίς την εφαρμογή εξομάλυνσης (smoothing)

Στην ενότητα αυτή θα δούμε τις ακολουθίες λέξεων μέσα από το πρίσμα γλωσσικών πιθανοτικών μοντέλων. Λέγοντας γλωσσικά πιθανοτικά μοντέλα εννοούμε τρόπους με τους οποίους ανατίθενται πιθανότητες σε ακολουθίες λέξεων. Με τα πιθανοτικά μοντέλα μπορεί να υπολογιστεί πιθανότητα που έχει κάποια πρόταση, όπως επίσης μπορεί να προβλεφθεί η πιθανότητα που έχει μια λέξη να είναι η επόμενη μέσα σε μία ακολουθία λέξεων.

Το πιο απλοϊκό μοντέλο για ακολουθίες λέξεων επιτρέπει σε κάθε λέξη να ακολουθείται από κάθε άλλη λέξη. Μιλώντας με πιθανότητες, κάνουμε την παραδοχή ότι κάθε λέξη έχει ίση πιθανότητα να ακολουθεί κάθε άλλη λέξη.

Σε ένα λίγο πιο πολύπλοκο μοντέλο, θα μπορούσαμε πάλι να επιτρέψουμε σε κάθε λέξη να ακολουθεί οποιαδήποτε άλλη λέξη, όμως η πιθανότητα εμφάνισης της λέξης θα σχετίζεται με την συχνότητα εμφάνισης της λέξης στο corpus. Για παράδειγμα στο Brown corpus, το οποίο περιέχει 1000000 λέξεις, η λέξη the εμφανίζεται 69971 φορές (7% των λέξεων) σε αντίθεση με την λέξη rabbit η οποία εμφανίζεται μόνο 11 φορές. Μπορούμε να χρησιμοποιήσουμε τις σχετικές συχνότητες των λέξεων για να προβλέψουμε ποια είναι η πιθανότητα κάθε λέξης να εμφανιστεί σε μία ήδη υπάρχουσα σειρά από λέξεις. Έτσι έχοντας την λέξη Anyhow, δίνουμε πιθανότητα 0.07 να είναι η επόμενη λέξη το the και πιθανότητα 0.00001 να είναι η επόμενη λέξη το rabbit. Όμως αυτή η θεώρηση, όπως θα διαπιστώσουμε με το επόμενο παράδειγμα, είναι κάπως προβληματική. Έστω ότι έχουμε την επόμενη ακολουθία λέξεων:

Just then, the white

Η παραπάνω σειρά λέξεων είναι λογικό να ακολουθείται από την λέξη rabbit και όχι από την λέξη the. Η προηγούμενη παρατήρηση μας υποδηλώνει ότι αντί να εξετάζουμε ανεξάρτητα τις σχετικές συχνότητες κάθε λέξης, θα ήταν καλύτερο να εξετάζουμε την δεσμευμένη πιθανότητα μίας λέξης δοθέντος ενός συνόλου από

προηγούμενες λέξεις. Αναφερόμενοι στο προηγούμενο παράδειγμα, δεν εξετάζουμε την πιθανότητα της λέξης rabbit ($P(rabbit)$), αλλά την πιθανότητα της λέξης rabbit δοθέντος της λέξης white ($P(rabbit|white)$) η οποία είναι πολύ μεγαλύτερη.

Έχοντας κάνει τις παραπάνω παρατηρήσεις, ας δούμε πως υπολογίζεται η πιθανότητα μιας αλληλουχίας λέξεων, την οποία συμβολίζουμε ως εξής: $w_1,\ w_2...,\ w_{n-1},\ w_n$ (w_1^n). Θεωρώντας ότι κάθε λέξη εμφανίζεται στην σωστή τοποθεσία σαν ένα ανεξάρτητο γεγονός, μπορούμε να παρουσιάσουμε την πιθανότητα της ακολουθίας των λέξεων ως εξής:

$$P(w_1,\ w_2...,\ w_{n-1},\ w_n)\ \ or\ \ P(w_1^n) \tag{2.3}$$

Χρησιμοποιώντας τον κανόνα της αλυσίδας μπορούμε να γράψουμε ότι:

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ ...\ P(w_n|w_1^{n-1}) = \\ &= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \end{aligned} \tag{2.4}$$

Ένα ερώτημα που ανακύπτει από τον παραπάνω τύπο είναι το εξής: Πώς μπορούν να υπολογιστούν πιθανότητες όπως το $P(w_n|w_1^{n-1})$; Η απάντηση είναι ότι δεν υπάρχει εύκολος τρόπος υπολογισμού της πιθανότητας που έχει μία λέξη, δοθέντος μίας μεγάλης ακολουθίας λέξεων οι οποίες προηγούνται. Για να λύσουμε αυτό το πρόβλημα κάνουμε την ακόλουθη προσέγγιση - απλοποίηση: Η πιθανότητα που έχει μία λέξη δοθέντος μίας ακολουθίας λέξεων προσεγγίζεται με την πιθανότητα της λέξης δοθέντος μόνο της προηγούμενης λέξης. Επομένως με αυτή την προσέγγιση (bigram model), η πιθανότητα $P(w_n|w_1^{n-1})$ προσεγγίζεται με την πιθανότητα $P(w_n|w_{n-1})$. Θέλοντας να δείξουμε τα παραπάνω με ένα παράδειγμα μπορούμε να γράψουμε ότι η πιθανότητα $P(horse|Yesterday\ i\ saw\ a\ white)$ προσεγγίζεται με την πιθανότητα $P(horse|white)$.

Η υπόθεση ότι η πιθανότητα μίας λέξης εξαρτάται μόνο από την προηγούμενη λέξη καλείται υπόθεση του Markov. Τα μοντέλα Markov είναι μια κλάση μοντέλων τα οποία θεωρούν ότι μπορούμε να προβλέψουμε την πιθανότητα που έχει μία λέξη, χωρίς να κοιτάξουμε πολύ βαθιά στο παρελθόν.

Γενικεύοντας το bigram μοντέλο, στο οποίο μια λέξη εξαρτάται μόνο από την

προηγούμενη λέξη, μπορούμε να περάσουμε στο N-gram μοντέλο, στο οποίο η λέξη εξαρτάται από τις N-1 προηγούμενες λέξεις. Το bigram καλείται ως πρώτης τάξης μοντέλο Markov (αφού "κοιτάζει" μια λέξη στο παρελθόν). Αντίστοιχα το trigram είναι ένα δεύτερης τάξης μοντέλο Markov (αφού "κοιτάζει" δύο λέξεις στο παρελθόν), και γενικά το N-gram είναι ένα N-1 τάξης μοντέλο Markov.

Η γενική εξίσωση προσέγγισης, με N-gram μοντέλο, της δεσμευμένης πιθανότητας που έχει μία λέξη, δοθέντος της ακολουθίας όλων των προηγούμενων λέξεων είναι η εξής:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1}) \tag{2.5}$$

Χρησιμοποιώντας την εξίσωση 2.5 και την εξίσωση 2.4 βλέπουμε ότι, για την περίπτωση του bigram μοντέλου, η πιθανότητα που έχει η πλήρης ακολουθία λέξεων προσεγγίζεται ως εξής:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1}) \tag{2.6}$$

Ένα σημαντικό πρόβλημα το οποίο διαφαίνεται από τον παραπάνω τύπο είναι το εξής: Εφόσον οι πιθανότητες είναι εξ ορισμού μικρότερες από το 1, όσο περισσότερες πιθανότητες πολλαπλασιάζουμε τόσο μικρότερο θα είναι το αποτέλεσμα με κίνδυνο να έχουμε προβλήματα numerical underflow. Η λύση στο προηγούμενο πρόβλημα είναι να παίρνουμε τον λογάριθμο κάθε πιθανότητας (logprob) και στην συνέχεια να αθροίζουμε τους λογαρίθμους. Πολλά προγράμματα, όπως το CMU toolkit το οποίο χρησιμοποιούμε στην εργασία μας, αποθηκεύουν και υπολογίζουν τις πιθανότητες των N-grams ως logprobs.

Ένα άλλο θέμα είναι το πως υπολογίζονται οι πιθανότητες των N-grams στην αρχή των προτάσεων. Για να γίνουμε περισσότερο σαφείς, έστω ότι θέλουμε να υπολογίσουμε την πιθανότητα της πρότασης: $w_1\ w_2\ w_3\ w_4$, χρησιμοποιώντας το bigram μοντέλο (τύπος 2.6). Είναι φανερό ότι η λέξη w_1 δεν έχει προϊστορία. Για να αντιμετωπίσουμε αυτό το πρόβλημα ενσωματώνουμε στην αρχή της πρότασης την ψευδολέξη start και ο υπολογισμός της πιθανότητας γίνεται ως εξής:

$$P(w_1^4) = P(w_1|start)\, P(w_2|w_1)\, P(w_3|w_2)\, P(w_4|w_3)$$

Παρόμοια στην περίπτωση που είχαμε trigram μοντέλο, θα χρησιμοποιούσαμε δύο ψευδολέξεις (start1, start2).

Πως όμως εκπαιδεύονται τα N-gram μοντέλα; Καταρχήν έχουμε τα δεδομένα από το training corpus. Από το corpus παίρνουμε τον αριθμό των εμφανίσεων ενός συγκεκριμένου N-gram. Στην συνέχεια διαιρούμε τον αριθμό αυτό, με το άθροισμα των εμφανίσεων όλων των N-grams, τα οποία έχουν τις ίδιες λέξεις πριν από την τελευταία, σε σχέση με το συγκεκριμένο N-gram που αναφέραμε προηγουμένως. Η προηγούμενη διαίρεση - κανονικοποίηση γίνεται ώστε το αποτέλεσμα (πιθανότητα του N-gram) να είναι αριθμός μεταξύ του 0 και του 1. Τα παραπάνω στην περίπτωση του bigram μοντέλου εκφράζονται ως εξής:

$$P(w_n|w_{n-1}) \;=\; \frac{C(w_{n-1}\; w_n)}{\sum_w C(w_{n-1}\; w)} \tag{2.7}$$

Το άθροισμα των εμφανίσεων των bigrams τα οποία αρχίζουν με την λέξη w_{n-1} είναι ίσο με τον αριθμό των εμφανίσεων της λέξης w_{n-1}. Οπότε μπορούμε να απλοποιήσουμε τον τύπο 2.7 ως εξής:

$$P(w_n|w_{n-1}) \;=\; \frac{C(w_{n-1}\; w_n)}{C(w_{n-1})} \tag{2.8}$$

Στην γενική περίπτωση των N-gram μοντέλων ο παραπάνω τύπος γίνεται:

$$P(w_n|w_{n-N+1}^{n-1}) \;=\; \frac{C(w_{n-N+1}^{n-1}\; w_n)}{C(w_{n-N+1}^{n-1})} \tag{2.9}$$

Στους παραπάνω τύπους η συνάρτηση C(.) επιστρέφει τον αριθμό των εμφανίσεων του ορίσματος της. Επίσης το κλάσμα του τύπου 2.9 καλείται σχετική συχνότητα (relative frequency). Η χρησιμοποίηση σχετικών συχνοτήτων για την εκτίμηση πιθανοτήτων είναι παράδειγμα μιας τεχνικής γνωστής ως Maximum Likelihood Estimation. Τα μοντέλα δηλαδή παραμετροποιούνται έτσι ώστε η πιθανότητα, του training set T δοθέντος του μοντέλου M ($P(T|M)$), να μεγιστοποιείται.

Κλείνοντας σημειώνουμε δύο σημαντικά θέματα για την "συμπεριφορά" των N-grams. Το πρώτο είναι η αυξανόμενη ακρίβεια των N-gram μοντέλων, αναφορικά με την αύξηση της τάξεως του μοντέλου. Το δεύτερο θέμα είναι η ισχυρή εξάρτηση που έχουν τα N-gram μοντέλα, αναφορικά με το training corpus. Η επιλογή του training corpus είναι ένα δύσκολο θέμα. Ένα N-gram μοντέλο εκπαιδεύεται με τα δεδομένα του corpus με στόχο να παραχθούν οι πιθανότητες των N-grams. Το corpus πρέπει να έχει σχεδιαστεί προσεκτικά, ώστε να προσφέρει ένα αντιπροσωπευτικό δείγμα των δεδομένων με τα οποία σχετίζεται η εφαρμογή. Εάν τα δεδομένα του corpus είναι υπερβολικά περιορισμένα στο πεδίο της εφαρμογής, ίσως οι πιθανότητες να είναι " προκατειλημμένες " και να μην μπορούν να γενικευθούν σε νέες προτάσεις. Από την άλλη πλευρά, αν τα δεδομένα είναι γενικά και δεν σχετίζονται αρκετά με το πεδίο ενδιαφέροντος, τότε οι πιθανότητες που θα παράγει το μοντέλο δεν θα είναι αντιπροσωπευτικές για την εφαρμογή, με αποτέλεσμα να υπάρχει μειωμένη απόδοση.

2.3 Smoothing

Κάθε corpus έχει περιορισμένο αριθμό προτάσεων. Έτσι είναι πιθανό κάποια N-grams να μην εμφανίζονται στο corpus, με αποτέλεσμα να τους ανατίθεται μηδενική πιθανότητα. Επιπλέον η χρησιμοποίηση μόνο σχετικών συχνοτήτων για την ανάθεση πιθανοτήτων στα N-grams, ίσως επιφέρει πολύ φτωχές εκτιμήσεις για τα N-grams που εμφανίζονται λίγες φορές στο corpus. Οι προηγούμενες παρατηρήσεις μας επισημαίνουν την ανάγκη επανεκτίμησης των πολύ μικρών και μηδενικών πιθανοτήτων και την ανάθεση μη μηδενικών τιμών. Η διαδικασία αυτή λέγεται smoothing και θα την αναλύσουμε στις επόμενες ενότητες.

2.3.1 Add-One smoothing

Το add-one smoothing είναι ένας απλούστατος αλγόριθμος smoothing ο οποίος στην πράξη δεν προσφέρει ικανοποιητικά αποτελέσματα. Ωστόσο είναι ένα καλό πρώτο βήμα για την κατανόηση των εννοιών που θα μας χρειαστούν στους επόμενους (πιο πολύπλοκους) αλγόριθμους. Ο αλγόριθμος αυτός προτείνει να προσθέσουμε (πριν την κανονικοποίηση τους σε πιθανότητες) το 1 στους αριθμούς των εμφανίσεων

καθενός N-gram type.

Για λόγους απλότητας ας θεωρήσουμε το add-one smoothing στην περίπτωση του unigram μοντέλου. Η maximum likelihood εκτίμηση της πιθανότητας των unigrams, χωρίς smoothing, είναι η εξής:

$$P(w_x) \;=\; \frac{C(w_x)}{\sum_i C(w_i)} \;=\; \frac{C(w_x)}{N} \tag{2.10}$$

Στον παραπάνω τύπο συμβολίζουμε με N, το συνολικό πλήθος των λέξεων.

Το smoothing βασίζεται στην προσαρμογή των αριθμήσεων. Συγκεκριμένα στην περίπτωση του add-one smoothing (στο unigram μοντέλο) η προσαρμογή αυτή γίνεται προσθέτοντας το 1 στο αριθμό των εμφανίσεων κάθε word type και πολλαπλασιάζοντας με ένα παράγοντα κανονικοποίησης ίσο με $\frac{N}{N+V}$. Στο προηγούμενο κλάσμα το N συμβολίζει το συνολικό πλήθος των λέξεων και το V συμβολίζει το πλήθος των διαφορετικών λέξεων (μέγεθος του λεξικού). Εφόσον προσθέτουμε το 1 στον αριθμό των εμφανίσεων κάθε word type, ο συνολικός αριθμός των λέξεων θα αυξηθεί κατά ποσότητα ίση με το μέγεθος του λεξικού. Σύμφωνα με τα παραπάνω, η προσαρμοσμένη αρίθμηση ορίζεται ως εξής:

$$c_i^* \;=\; (c_i \;+\; 1)\,\frac{N}{N+V} \tag{2.11}$$

Οι παραπάνω προσαρμοσμένες αριθμήσεις μπορούν να μετατραπούν σε πιθανότητες (p_i^*) αν τις κανονικοποιήσουμε με το N:

$$p_i^* \;=\; \frac{c_i \;+\; 1}{N \;+\; V} \tag{2.12}$$

Χρησιμοποιώντας τους τύπους 2.12 και 2.8, μπορούμε να γράψουμε ότι οι add-one smoothed πιθανότητες των bigrams δίδονται από τον ακόλουθο τύπο:

$$p_i^* \;=\; \frac{C(w_{n-1}w_n) \;+\; 1}{C(w_{n-1}) \;+\; V} \tag{2.13}$$

Μία διαφορετική οπτική γωνία με την οποία μπορεί να δει κάποιος το smoothing είναι η εξής: Ένας αλγόριθμος smoothing κάνει " εκπτώσεις " (discounting) σε

κάποιους μη μηδενικούς αριθμούς εμφανίσεων N-grams, ώστε να αποταμιεύσει ορισμένη μάζα πιθανότητας την οποία θα διαμοιράσει στα N-grams με μηδενικό αριθμό εμφανίσεων. Το κλάσμα των προσαρμοσμένων αριθμήσεων (c*) προς τον αριθμό των αυθεντικών αριθμήσεων είναι το discount ratio (d_c):

$$d_c = \frac{c^*}{c} \tag{2.14}$$

Η επιλογή πρόσθεσης του 1 στους αριθμούς των εμφανίσεων των N-grams είναι αυθαίρετη. Αυτό ίσως να δημιουργήσει προβλήματα διότι μπορεί να μετατοπιστεί υπερβολικά πολύ μάζα πιθανότητας προς τα N-grams με μηδενικό αριθμό εμφανίσεων. Θα μπορούσαμε να αντιμετωπίσουμε το πρόβλημα αυτό αλλάζοντας τον αριθμό που προσθέτουμε (αντικαθιστώντας το 1 με μία μικρότερη τιμή), δηλαδή θα μπορούσαμε να έχουμε add-one-half smoothing, add-one-thousandth smoothing, κ.τ.λ.

Κλείνοντας να σημειώσουμε ότι το add-one smoothing είναι μια μέθοδος smoothing χωρίς αξιόλογα αποτελέσματα. Έχει επισημανθεί από τους Gale και Church ότι οι διακυμάνσεις των αριθμήσεων που προκύπτουν από το add-one smoothing, είναι χειρότερες από αυτές που παράγονται από την unsmoothed maximum likelihood εκτίμηση.

2.3.2 Witten-Bell Discounting

Όπως προαναφέραμε στην προηγούμενη ενότητα, ο αλγόριθμος του add-one smoothing δεν προσφέρει ικανοποιητικά αποτελέσματα. Ένας καλύτερος (όμως περισσότερο πολύπλοκος) αλγόριθμος smoothing είναι το Witten-Bell discounting. Ο αλγόριθμος αυτός βασίζεται σε μια έξυπνη παρατήρηση αναφορικά με τα γεγονότα μηδενικής συχνότητας. Αν θεωρήσουμε ένα μηδενικής συχνότητας N-gram ως ένα N-gram το οποίο δεν έχει συμβεί ακόμα, τότε όταν συμβεί θα είναι η πρώτη φορά που θα το δούμε. Έτσι η πιθανότητα να συμβεί ένα N-gram μηδενικής συχνότητας μπορεί να μοντελοποιηθεί από την πιθανότητα που έχει ένα N-gram να συμβεί για πρώτη φορά.

Πως όμως μπορούμε να υπολογίσουμε την πιθανότητα που έχει ένα N-gram να συμβεί για πρώτη φορά; Απλά μετράμε τις φορές που βλέπουμε N-grams για πρώτη φορά στο training corpus. Ο αριθμός των N-grams που βλέπουμε για πρώτη φορά είναι ίσος με τον αριθμό των διαφορετικών N-grams (N-gram types) που υπάρχουν στο training corpus.

Έχοντας υπόψιν τα παραπάνω, μπορούμε να γράψουμε ότι η συνολική μάζα πιθανότητας η οποία θα διαμοιραστεί στα N-grams μηδενικής συχνότητας είναι η εξής:

$$\sum_{i:c_i=0} p_i^* \;=\; \frac{T}{N \;+\; T} \tag{2.15}$$

Στον παραπάνω τύπο το T συμβολίζει το πλήθος των διαφορετικών N-grams που έχουμε ήδη δει και το N το συνολικό πλήθος των N-grams.

Ο τύπος 2.15, όπως προαναφέραμε, δίνει την συνολική μάζα πιθανότητας η οποία θα διαμοιραστεί στα N-grams που δεν έχουμε δει. Η απλούστερη προσέγγιση είναι να κατανείμουμε ισότιμα την ποσότητα αυτή ανάμεσα στα N-grams μηδενικής συχνότητας. Έστω Z ο αριθμός των N-grams μηδενικής συχνότητας:

$$Z \;=\; \sum_{i:c_i=0} 1 \tag{2.16}$$

Τότε η ποσότητα πιθανότητας που αναλογεί σε ένα N-gram μηδενικής συχνότητας είναι:

$$p_i^* \;=\; \frac{T}{Z\,(N+T)} \tag{2.17}$$

Ένα ερώτημα που ανακύπτει σε αυτό το σημείο είναι το εξής: Από που θα βρεθεί η ποσότητα πιθανότητας (τύπος 2.15) η οποία θα διαμοιραστεί στα N-grams μηδενικής συχνότητας; Η απάντηση είναι ότι θα υποβαθμίσουμε (discount) τις πιθανότητες των N-grams μη μηδενικής συχνότητας ως εξής:

$$p_i^* \;=\; \frac{c_i}{N+T} \;\; if(c_i > 0) \tag{2.18}$$

Επίσης μπορούμε να γράψουμε ότι προσαρμοσμένες αριθμήσεις (smoothed counts) είναι οι παρακάτω:

$$c_i^* = \begin{cases} \frac{T}{Z}\frac{N}{N+T} & if\ c_i = 0 \\ c_i \frac{N}{N+T} & if\ c_i > 0 \end{cases} \tag{2.19}$$

Σε αυτό το σημείο αξίζει να δούμε το Witten-Bell discounting στην περίπτωση των bigrams ώστε να δούμε την μεγάλη διαφορά που υπάρχει με το add-one smoothing. Η διαφορά εντοπίζεται στο ότι οι αριθμήσεις μας εξαρτώνται από κάποια ιστορία. Για να υπολογιστεί η πιθανότητα ενός bigram $w_{n-1}w_{n-2}$ το οποίο δεν έχουμε δει, χρησιμοποιείται η πιθανότητα του να δούμε ένα νέο bigram το οποίο ξεκινάει με το w_{n-1}. Αυτό σημαίνει ότι η εκτίμηση της πιθανότητας που θα έχει ένα bigram που δεν έχουμε δει, θα βασίζεται σε ιστορία μιας λέξης. Λέξεις οι οποίες υπάρχουν σε λίγα bigrams θα παρέχουν φτωχότερες εκτιμήσεις από τις πιο πολλά υποσχόμενες λέξεις.

Σύμφωνα με τα προηγούμενα και τον τύπο 2.15, η συνολική ποσότητα της μάζας πιθανότητας η οποία θα διατεθεί στα bigrams $w_x w_i$ που δεν έχουμε δει υπολογίζεται ως εξής:

$$\sum_{i:c(w_x w_i)=0} p^*(w_i|w_x) \ = \ \frac{T(w_x)}{N(w_x) \ + \ T(w_x)} \tag{2.20}$$

Στον παραπάνω τύπο το $T(w_x)$ είναι το πλήθος των διαφορετικών bigrams τα οποία έχουμε ήδη δει και ξεκινάνε με το w_x. Το $N(w_x)$ είναι το πλήθος όλων των bigrams που ξεκινάνε με το w_x.

Κατανέμοντας την ποσότητα της μάζας πιθανότητας του τύπου 2.20 στα bigrams που δεν έχουμε δει παίρνουμε τον ακόλουθο τύπο:

$$p^*(w_i|w_{i-1}) \ = \ \frac{T(w_{i-1})}{Z(w_{i-1})\,(N \ + \ T(w_{i-1}))} \quad if(c_{w_{i-1}w_i} = 0) \tag{2.21}$$

Το $Z(w_{i-1})$, στον παραπάνω τύπο, είναι το συνολικό πλήθος των bigrams τα οποία ξεκινάνε με την λέξη w_{i-1} και έχουν μηδενικό αριθμό εμφανίσεων.

Όσον αφορά τα bigrams με μη μηδενικό αριθμό εμφανίσεων, παραμετροποιούμε το T σύμφωνα με την ιστορία:

$$p^*(w_i|w_x) \;=\; \frac{c(w_x w_i)}{c(w_x) \;+\; T(w_x)} \tag{2.22}$$

2.3.3 Good-Turing Discounting

Σε αυτή την ενότητα θα ασχοληθούμε με το Good-Turing discounting το οποίο είναι ένας πιο πολύπλοκος αλγόριθμος smoothing σε σχέση με το Witten-Bell discounting.

Η βασική ιδέα στον αλγόριθμο Good-Turing είναι η επανεκτίμηση της ποσότητας μάζας πιθανότητας, η οποία διατίθεται στα N-grams με μηδενικό ή μικρό αριθμό εμφανίσεων, χρησιμοποιώντας τον αριθμό των N-grams που έχουν μεγαλύτερους αριθμούς εμφανίσεων. Ορίζουμε ως N_c, τον αριθμό των N-grams τα οποία εμφανίζονται c φορές.

Ένα διαφορετικό είδος discounting σε σχέση με αυτό που είδαμε στην προηγούμενη ενότητα είναι το "non-conditional discounting". Συγκεκριμένα στην προηγούμενη ενότητα (Witten-Bell discounting) οι smoothed πιθανότητες των bigrams ήταν δεσμευμένες πιθανότητες λέξεων, εξαρτώμενων από την προηγούμενη λέξη. Όμως μπορούμε να θεωρήσουμε το bigram σαν αυτόνομη μονάδα (αγνοώντας το γεγονός ότι αποτελείται από δύο λέξεις), κάνοντας discounting όχι στην δεσμευμένη πιθανότητα $p(w_i|w_x)$ αλλά στην από κοινού πιθανότητα (joint probability) $p(w_x w_i)$. Η θεώρηση αυτή (non-conditional discounting) χρησιμοποιείται στο Good-Turing discounting.

Με βάση την παραπάνω θεώρηση για την εξομάλυνση (smoothing) της joint

probability των bigrams, N_c είναι ο αριθμός των bigrams b που έχουν αριθμό εμφανίσεων ίσο με c. Οπότε έχουμε:

$$N_c = \sum_{b:c(b)=c} 1 \tag{2.23}$$

Στην Good-Turing εκτίμηση οι προσαρμοσμένες αριθμήσεις (smoothed counts) δίνονται από τον ακόλουθο τύπο:

$$c^* = (c+1)\,\frac{N_{c+1}}{N_c} \tag{2.24}$$

Παραδείγματος χάριν, στην περίπτωση των bigrams με μηδενικό αριθμό εμφανίσεων έχουμε:

$$c_0^* = (0+1)\,\frac{N_1}{N_0}$$

Από τον τον παραπάνω τύπο φαίνεται το σκεπτικό που είχαμε δει και στο Witten-Bell discounting, δηλαδή την χρησιμοποίηση των πραγμάτων που έχουμε δει μία φορά στην εκτίμηση των πραγμάτων που δεν έχουμε δει ακόμα.

Σε αυτό το σημείο τίθεται στον αναγνώστη το εξής ερώτημα: Πως ξέρουμε τον αριθμό των bigrams που δεν έχουμε δει (N_0); Η απάντηση σε αυτό το ερώτημα είναι ότι ξέροντας το μέγεθος του λεξικού (V), ο συνολικός αριθμός των bigrams είναι V^2. Έτσι μπορούμε να πούμε ότι το N_0 προκύπτει αν αφαιρέσουμε από το V^2 τον αριθμό των bigrams που έχουμε δει.

Στην πράξη οι προσαρμογή των αριθμήσεων (c^*) δεν γίνεται για όλες τις αριθμήσεις c. Για κάποιο κατώφλι (threshold) k, ο τύπος 2.24 γίνεται:

$$c^* = c \;\; for\; c > k \tag{2.25}$$

$$c^* = \frac{(c+1)\frac{N_{c+1}}{N_c} - c\frac{(k+1)N_{k+1}}{N_1}}{1 \; - \; \frac{(k+1)N_{k+1}}{N_1}} \; , \; for \; 1 \leq c \leq k. \tag{2.26}$$

2.4 Backoff

Στην προηγούμενη ενότητα ασχοληθήκαμε με διάφορους αλγόριθμους smoothing οι οποίοι μας βοήθησαν να αντιμετωπίσουμε το πρόβλημα των N-grams μηδενικής συχνότητας. Υπάρχει όμως και ένα επιπλέον θέμα με το οποίο πρέπει να ασχοληθούμε. Εάν δεν υπάρχουν παραδείγματα ενός trigram $w_{n-2}w_{n-1}w_n$ για να μας βοηθήσουν να υπολογίσουμε την πιθανότητα $P(w_n|w_{n-2}w_{n-1})$, μπορούμε να την εκτιμήσουμε χρησιμοποιώντας την πιθανότητα του bigram $w_{n-1}w_n$ ($P(w_n|w_{n-1})$). Παρόμοια αν δεν υπάρχουν παραδείγματα για να υπολογίσουμε την πιθανότητα $P(w_n|w_{n-1})$, μπορούμε να χρησιμοποιήσουμε την πιθανότητα που έχει το unigram w_n ($P(w_n)$).

Ένας τρόπος να χρησιμοποιήσουμε αυτήν την ιεραρχία των N-grams ώστε να χτίσουμε ένα N-gram μοντέλο είναι το backoff. Στο backoff μοντέλο χτίζουμε ένα N-gram μοντέλο βασιζόμενοι σε ένα (N-1)-gram μοντέλο. Κάτι που πρέπει να τονιστεί εδώ είναι ότι στο backoff μοντέλο περνάμε σε ένα μικρότερης τάξης N-gram μόνο όταν δεν υπάρχουν παραδείγματα ενός N-gram μεγαλύτερης τάξης.

Σύμφωνα με τα παραπάνω στην περίπτωση του trigram το backoff μοντέλο περιγράφεται ως εξής:

$$\widehat{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} P(w_i|w_{i-2}w_{i-1}), & if \; C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 P(w_i|w_{i-1}), & if \; C(w_{i-2}w_{i-1}w_i) = 0 \\ & and \; C(w_{i-1}w_i) > 0 \\ \alpha_2 P(w_i), & otherwise \end{cases} \tag{2.27}$$

Τα βάρη α_1, α_2 χρησιμοποιούνται στον παραπάνω τύπο, ώστε το αποτέλεσμα της εξίσωσης να έχει έγκυρη τιμή. Στην γενική περίπτωση η αναδρομική εξίσωση του backoff είναι:

$$\begin{aligned}\widehat{P}(w_n|w_{n-N+1}^{n-1}) &= \tilde{P}(w_n|w_{n-N+1}^{n-1}) \\ &\quad + \theta(P(w_n|w_{n-N+1}^{n-1}))\alpha\widehat{P}(w_n|w_{n-N+2}^{n-1})\end{aligned} \tag{2.28}$$

Στον παραπάνω τύπο χρησιμοποιούμε την συνάρτηση $\theta(.)$, την οποία ορίζουμε παρακάτω, ώστε να απαιτήσουμε από το backoff μοντέλο να επιλέγει χαμηλότερης τάξης μοντέλο, μόνο όταν το υψηλότερης τάξης μοντέλο δίνει μηδενική πιθανότητα.

$$\theta(x) = \begin{cases} 1, & if\ x = 0 \\ 0, & otherwise \end{cases} \tag{2.29}$$

Τέλος να σημειώσουμε ότι κάθε πιθανότητα $P(.)$ στον τύπο 2.28 προέρχεται από maximum likelihood εκτίμηση.

Συνδυάζοντας backoff και discounting: Στις προηγούμενες ενότητες, στις οποίες χρησιμοποιήσαμε το discounting για να βρούμε την ποσότητα της μάζας πιθανότητας που διατίθεται στα γεγονότα μηδενικής εμφάνισης, θεωρήσαμε ότι τα γεγονότα αυτά ήταν ισοδύναμα. Το αποτέλεσμα της θεώρησης αυτής ήταν ο ισότιμος διαμερισμός της ποσότητας μάζας πιθανότητας μεταξύ των γεγονότων μηδενικής εμφάνισης. Μια έξυπνη ιδέα είναι να συνδυάσουμε το discounting με τον αλγόριθμο του backoff, ώστε να αναθέτουμε πιο αποδοτικά τις πιθανότητες στα γεγονότα αυτά. Έτσι θα χρησιμοποιήσουμε τον αλγόριθμο του discounting ώστε να ξέρουμε πόση ποσότητα μάζας πιθανότητας διατίθεται στα γεγονότα μηδενικής εμφάνισης, και τον αλγόριθμο του backoff ώστε να κατανείμουμε την ποσότητα αυτή με έναν έξυπνο τρόπο.

Είναι σημαντικό να κατανοήσει ο αναγνώστης την σημασία των τιμών του α στην εξίσωση 2.28. Οι τιμές που παίρνει το α είναι βάρη τέτοια ώστε το αποτέλεσμα της εξίσωσης 2.28 να είναι πραγματική πιθανότητα (να έχει έγκυρη τιμή). Αν δεν χρησιμοποιούσαμε τα βάρη αυτά το αποτέλεσμα θα ήταν μεγαλύτερο του 1.

Χρησιμοποιώντας σχετικές συχνότητες (relative frequencies), εάν αθροίσουμε την πιθανότητα μιας λέξης w_n μέσα σε όλα τα δυνατά N-gram contexts, τότε το

αποτέλεσμα είναι ίσο με την μονάδα. Έτσι μπορούμε να γράψουμε ότι:

$$\sum_{i,j} P(w_n|w_i w_j) \;=\; 1$$

Σε αυτήν την περίπτωση, εάν κάνουμε back off σε ένα χαμηλότερης τάξης μοντέλο όταν η πιθανότητα είναι μηδέν, θα προσθέσουμε επιπλέον μάζα πιθανότητας στην εξίσωση και το αποτέλεσμα θα είναι μεγαλύτερο του 1. Έτσι είναι φανερό ότι πρέπει να εφαρμοστεί ένας discounting αλγόριθμος στο backoff μοντέλο. Έτσι η τελική μορφή της εξίσωσης 2.28 είναι:

$$\begin{aligned} \widehat{P}(w_n|w_{n-N+1}^{n-1}) \;&= \tilde{P}(w_n|w_{n-N+1}^{n-1}) \\ &\quad + \theta(P(w_n|w_{n-N+1}^{n-1})).\alpha(w_{n-N+1}^{n-1})\widehat{P}(w_n|w_{n-N+2}^{n-1}) \end{aligned} \tag{2.30}$$

Στον παραπάνω τύπο, το $\tilde{P}(.)$ αφορά τις discounted maximum likelihood εκτιμήσεις των πιθανοτήτων:

$$\tilde{P}(w_n|w_{n-N+1}^{n-1}) = \frac{c^*(w_{n-N+1}^{n})}{c(w_1^{n-N+1})} \tag{2.31}$$

Η συνάρτηση α(.) αντιπροσωπεύει την ποσότητα της μάζας πιθανότητας η οποία πρέπει κατανεμηθεί από ένα N-gram σε ένα (N-1)-gram και δίνεται από το ακόλουθο τύπο:

$$a(w_n|w_{n-N+1}^{n-1}) = \frac{1 - \sum_{w_n:c(w_{n-N+1}^{n-1})>0} \tilde{P}(w_n|w_{n-N+1}^{n-1})}{1 - \sum_{w_n:c(w_{n-N+1}^{n-1})>0} \tilde{P}(w_n|w_{n-N+2}^{n-1})} \tag{2.32}$$

Κλείνοντας παραθέτουμε το backoff μοντέλο στην περίπτωση του trigram:

$$\widehat{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} \tilde{P}(w_i|w_{i-2}w_{i-1}), & if\ C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha(w_{n-2}^{n-1})\tilde{P}(w_i|w_{i-1}), & if\ C(w_{i-2}w_{i-1}w_i) = 0 \\ & and\ C(w_{i-1}w_i) > 0 \\ \alpha(w_{n-1})\tilde{P}(w_i), & otherwise \end{cases} \tag{2.33}$$

Κεφάλαιο 3

Το μοντέλο διανυσματικού χώρου

Στο μοντέλο διανυσματικού χώρου (Vector Space Model - VSM), κάθε κείμενο (όπως και κάθε ερώτηση πάνω στην συλλογή των κειμένων) αναπαριστάται από ένα διάνυσμα όρων (terms), του οποίου το μήκος ισούται με το πλήθος των μοναδικών γνωρισμάτων των κειμένων στην συλλογή. Κάθε στοιχείο του διανύσματος έχει ένα βάρος το οποίο δείχνει πόσο σημαντικός είναι ο όρος στον χαρακτηρισμό του κειμένου. Η φάση της εξαγωγής των όρων που χαρακτηρίζουν ένα κείμενο καλείται ευρετηρίαση κειμένου (document indexing).

Η συλλογή των κειμένων αναπαριστάται από τον πίνακα όρων-κειμένων (term document matrix), οι γραμμές του οποίου αντιπροσωπεύουν τους όρους και οι στήλες τα κείμενα. Στον πίνακα 3.1 φαίνεται ο πίνακας μιας συλλογής κειμένων η οποία έχει N μοναδικούς όρους και D κείμενα. Στον πίνακα 3.2 φαίνεται το διάνυσμα μίας ερώτησης πάνω στην συλλογή των κειμένων.

Όπως φαίνεται και από τον συμβολισμό στον πίνακα 3.1, κάθε όρος έχει ένα βάρος σε κάθε κείμενο. Έτσι το βάρος w_{ij}, φανερώνει την σημαντικότητα του όρου i στον χαρακτηρισμό του κειμένου j.

	document 1	document 2	...	document D
term 1	w_{11}	w_{12}	...	w_{1D}
term 2	w_{21}	w_{22}	...	w_{2D}
⋮	⋮	⋮	⋱	⋮
term N	w_{N1}	w_{N2}	...	w_{ND}

Πίνακας 3.1: Πίνακας όρων-κειμένων

	query
term 1	w_{11}^q
term 2	w_{21}^q
⋮	⋮
term N	w_{N1}^q

Πίνακας 3.2: Διάνυσμα ερώτησης πάνω στην συλλογή των κειμένων

3.1 Σχήμα ανάθεσης βαρών Tf-Idf

Υπάρχουν πολλά σχήματα ανάθεσης βαρών. Εδώ θα αναλύσουμε το tf-idf σχήμα ανάθεσης βαρών, το οποίο είναι πολύ διαδεδομένο και αποδοτικό. Έστω ένας όρος i. Η συχνότητα του όρου i στο κείμενο j (f_{ij}), είναι ο αριθμός των εμφανίσεων του όρου i στο κείμενο j. Κανονικοποιώντας αυτή την συχνότητα με την συχνότητα του όρου ο οποίος έχει την μέγιστη συχνότητα από όλους τους όρους δεικτοδότησης που εμφανίζονται μέσα στο κείμενο, μπορούμε να γράψουμε ότι η κανονικοποιημένη συχνότητα του όρου (term frequency) είναι:

$$tf_{ij} = \frac{f_{ij}}{\max_l f_{lj}} \tag{3.1}$$

Βλέπουμε ότι εάν $f_{ij} = 0$, τότε $tf_{ij} = 0$. Επίσης όσο πιο μεγάλο είναι το f_{ij}, τόσο καλύτερα περιγράφεται το κείμενο j από τον όρο i. Το επόμενο βήμα είναι η εξέταση της επίδρασης του όρου, όχι μόνο μέσα σε ένα κείμενο, αλλά σε όλη τη συλλογή. Διαισθητικά καταλαβαίνουμε ότι ένας όρος που εμφανίζεται σε λίγα

κείμενα, είναι καταλληλότερος για να διαχωρίσει τα κείμενα της συλλογής, από ένα όρο που εμφανίζεται σε όλα ή στα περισσότερα κείμενα της συλλογής. Ένας τρόπος για να εκφραστεί η παραπάνω παρατήρηση, είναι η αντίστροφη συχνότητα εμφάνισης κειμένου (inverse document frequency), την όποια ορίζουμε ως εξής:

$$idf_i = \log \frac{1 + D}{n_i} \tag{3.2}$$

Στον παραπάνω τύπο, συμβολίζουμε με D το πλήθος των κειμένων της συλλογής και με n_i τον αριθμό των κειμένων της συλλογής στα οποία εμφανίζεται ο όρος i. Όπως καταλαβαίνουμε, πριμοδοτούνται οι όροι οι οποίοι εμφανίζονται σε λίγα κείμενα, αφού χαρακτηρίζουν καλύτερα την κλάση των εγγράφων. Έχοντας ορίσει τα term frequency και inverse document frequency, μπορούμε να ορίσουμε το σχήμα ανάθεσης βαρών tf-idf ως εξής:

$$w_{ij} = tf_{ij}\ idf_i \tag{3.3}$$

Το w_{ij} είναι το βάρος του όρου i στο κείμενο j.

3.2 Εύρεση ομοιότητας των κειμένων με ερωτήσεις πάνω στην συλλογή

Όπως αναφέραμε και προηγουμένως, κάθε κείμενο d_j αναπαριστάται από ένα διάνυσμα $\mathbf{d_j}(w_{1j}, w_{2j}, ..., w_{Nj})$. Επίσης κάθε ερώτηση q αναπαριστάται από ένα διάνυσμα $\mathbf{q}(w^q_{11}, w^q_{21}, ..., w^q_{N1})$. Θυμίζουμε ότι N είναι το πλήθος των μοναδικών όρων δεικτοδότησης στην συλλογή κειμένων.

Η ομοιότητα μεταξύ του διανύσματος ενός κειμένου d_j και του διανύσματος μίας

ερώτησης q, βρίσκεται από το συνημίτονο της μεταξύ τους γωνίας (cosine similarity):

$$sim(\mathbf{d_j}, \mathbf{q}) = \frac{\mathbf{d_j}.\mathbf{q}}{|\mathbf{d_j}||\mathbf{q}|} = \frac{\sum_{1 \leq i \leq N} w_{ij} w_{i1}^q}{\sqrt{\sum_{1 \leq i \leq N} w_{ij}^2} \sqrt{\sum_{1 \leq i \leq N} (w_{i1}^q)^2}} \tag{3.4}$$

Ας δείξουμε όμως με ένα παράδειγμα, τον τρόπο με τον οποίο υπολογίζεται η ομοιότητα των κειμένων της συλλογής με τις ερωτήσεις πάνω στην συλλογή. Έστω ότι η συλλογή αποτελείται από τα εξής κείμενα:

d1: Shipment of gold damaged in a fire.
d2: Delivery of silver arrived in a silver truck.
d3: Shipment of gold arrived in a truck.

Από την συλλογή αφαιρούνται τα σημεία στίξης και όλοι οι χαρακτήρες γίνονται μικροί. Η ερώτηση q πάνω στην συλλογή είναι η εξής: "gold silver truck". Ο τρόπος υπολογισμού του πίνακα όρων-κειμένων γίνεται σύμφωνα με το σχήμα ανάθεσης βαρών που αναλύθηκε στην ενότητα 3.1. Το βάρος που ανατίθεται σε κάθε όρο της ερώτησης q, ισούται απλά με το πλήθος των εμφανίσεων του όρου στο κείμενο της ερώτησης. Στον πίνακα 3.3 φαίνεται αναλυτικά η διαδικασία.

Οι ομοιότητες των διανυσμάτων των κειμένων με το διάνυσμα της ερώτησης, υπολογίζονται σύμφωνα με το συνημίτονο της γωνίας μεταξύ των διανυσμάτων (cosine similarity (τύπος 3.4)). Συγκεκριμένα έχουμε:
Τα μήκη των κειμένων υπολογίζονται ως εξής:

$$|\mathbf{d1}| = \sqrt{0.12^2 + 0.6^2 + 0.6^2 + 0.3^2 + 0.12^2 + 0.12^2 + 0.3^2} = 0.97$$

$$|\mathbf{d2}| = \sqrt{0.06^2 + 0.15^2 + 0.3^2 + 0.06^2 + 0.06^2 + 0.6^2 + 0.15^2} = 0.71$$

$$|\mathbf{d3}| = \sqrt{0.12^2 + 0.3^2 + 0.3^2 + 0.12^2 + 0.12^2 + 0.3^2 + 0.3^2} = 0.63$$

	$tf_{ij} = \frac{f_{ij}}{\max_l f_{lj}}$				$w_{ij} = tf_{ij}\, idf_i$			
Terms	d1	d2	d3	$idf_i = \log \frac{1+D}{n_i}$	d1	d2	d3	q
a	1	0.5	1	$log(4/3) = 0.12$	0.12	0.06	0.12	0
arrived	0	0.5	1	$log(4/2) = 0.3$	0	0.15	0.3	0
damaged	1	0	0	$log(4/1) = 0.6$	0.6	0	0	0
delivery	0	0.5	0	$log(4/1) = 0.6$	0	0.3	0	0
fire	1	0	0	$log(4/1) = 0.6$	0.6	0	0	0
gold	1	0	1	$log(4/2) = 0.3$	0.3	0	0.3	1
in	1	0.5	1	$log(4/3) = 0.12$	0.12	0.06	0.12	0
of	1	0.5	1	$log(4/3) = 0.12$	0.12	0.06	0.12	0
shipment	1	0	1	$log(4/2) = 0.3$	0.3	0	0.3	0
silver	0	1	0	$log(4/1) = 0.6$	0	0.6	0	1
truck	0	0.5	1	$log(4/2) = 0.3$	0	0.15	0.3	1

Πίνακας 3.3: Παράδειγμα υπολογισμού των βαρών στον πίνακα όρων-κειμένων και στο διάνυσμα της ερώτησης

Το μήκος της ερώτησης είναι:

$$|\mathbf{q}| = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3} = 1.73$$

Σύμφωνα με τα παραπάνω, οι ομοιότητες υπολογίζονται ως εξής:

$$sim(\mathbf{d1}, \mathbf{q}) = \frac{0.3}{0.97\ .\ 1.73} = 0.179$$

$$sim(\mathbf{d2}, \mathbf{q}) = \frac{0.6+0.15}{0.71\ .\ 1.73} = 0.61$$

$$sim(\mathbf{d3}, \mathbf{q}) = \frac{0.3+0.3}{0.63\ .\ 1.73} = 0.55$$

Όπως φαίνεται από τα αποτελέσματα, η κατάταξη των κειμένων αναφορικά με την ομοιότητα τους με την ερώτηση είναι: $d2 > d3 > d1$

Αναφορές

Ο αναγνώστης μπορεί να ανατρέξει στις πηγές [11], [1], για επιπλέον πληροφορίες σχετικά με το μοντέλο διανυσματικού χώρου.

Κεφάλαιο 4

Λανθάνουσα σημασιολογική ανάλυση

Με το μοντέλο διανυσματικού χώρου, το οποίο περιγράψαμε στην προηγούμενη ενότητα, δύο διαφορετικές λέξεις έχουν μηδενική ομοιότητα. Ο λόγος είναι ότι αντιστοιχίζονται σε διαφορετική διάσταση στον διανυσματικό χώρο. Βέβαια αυτό δεν είναι πάντα επιθυμητό, λόγω του φαινομένου της συνωνυμίας. Θα θέλαμε δύο λέξεις οι οποίες έχουν παραπλήσια ή ίδια σημασία να μην έχουν μηδενική ομοιότητα. Με την λανθάνουσα σημασιολογική ανάλυση (Latent Semantic Analysis - LSA) αντιμετωπίζεται αυτό το πρόβλημα. Η ιδέα στην λανθάνουσα σημασιολογική ανάλυση είναι ότι δεν χρειάζονται όλες οι ορθογώνιες διαστάσεις, λόγω των σημασιολογικών ομοιοτήτων που παρουσιάζουν αρκετές λέξεις.

Έστω A ο πίνακας όρων-κειμένων του μοντέλου διανυσματικού χώρου, όπως τον περιγράψαμε στην προηγούμενη ενότητα. Στην λανθάνουσα σημασιολογική ανάλυση ο πίνακας A διασπάται σε γινόμενο τριών πινάκων με την τεχνική της διάσπασης ιδιοτιμών (Singular Value Decomposition - SVD) ως εξής:

$$A = U\,S\,V^T \tag{4.1}$$

, όπου:

$$A: \begin{bmatrix} w_{11} & w_{12} & ... & w_{1D} \\ w_{21} & w_{22} & ... & w_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & ... & w_{ND} \end{bmatrix} \quad U: \begin{bmatrix} u_{11} & u_{12} & ... & u_{1r} \\ u_{21} & u_{22} & ... & u_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & ... & u_{Nr} \end{bmatrix}$$

$$S: \begin{bmatrix} s_{11} & s_{12} & ... & s_{1r} \\ s_{21} & s_{22} & ... & s_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{r1} & s_{r2} & ... & s_{rr} \end{bmatrix} \quad V: \begin{bmatrix} v_{11} & v_{12} & ... & v_{1r} \\ v_{21} & v_{22} & ... & v_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ v_{D1} & v_{D2} & ... & v_{Dr} \end{bmatrix}$$

Θυμίζουμε ότι στον πίνακα όρων-κειμένων A, μεγέθους N x D, οι γραμμές αντιπροσωπεύουν τους όρους (οι οποίοι έχουν πλήθος N) και οι στήλες αντιπροσωπεύουν τα κείμενα (τα οποία έχουν πλήθος D).

Το εσωτερικό γινόμενο δύο γραμμών του πίνακα A, δείχνει κατά πόσο σχετίζονται μεταξύ τους οι όροι που αφορούν τις δύο γραμμές. Οι όροι που συνυπάρχουν σε πολλά κείμενα (ανάλογα βέβαια και με το βάρος των όρων στα κείμενα) θα παρουσιάζουν μεγάλη συσχέτιση-ομοιότητα. Ο πίνακας AA^T είναι ο πίνακας ομοιότητας όρων, είναι συμμετρικός και περιλαμβάνει όλα τα εσωτερικά γινόμενα που εκφράζουν την σχέση μεταξύ όρων. Το στοιχείο (i,j) του πίνακα αυτού, είναι η ομοιότητα του όρου i με τον όρο j.

Αντιστοίχως, το εσωτερικό γινόμενο δύο στηλών του πίνακα A, δείχνει κατά πόσο σχετίζονται μεταξύ τους τα κείμενα που αφορούν τις δύο στήλες. Τα κείμενα στα οποία υπάρχουν πολλοί ίδιοι όροι (ανάλογα βέβαια και με το βάρος τους) παρουσιάζουν μεγάλη συσχέτιση-ομοιότητα. Ο πίνακας A^TA είναι ο πίνακας ομοιότητας κειμένων, είναι συμμετρικός και περιλαμβάνει όλα τα εσωτερικά γινόμενα που εκφράζουν την σχέση μεταξύ κειμένων. Το στοιχείο (i,j) του πίνακα αυτού, είναι η ομοιότητα του κειμένου i με το κείμενο j.

Ο U είναι ένας ορθογώνιος πίνακας μεγέθους N x r, ο οποίος περιέχει τα ιδιοδιανύσματα του AA^T. Ο S είναι διαγώνιος πίνακας διάστασης r x r, όπου r είναι η τάξη του πίνακα A. Τα διαγώνια στοιχεία του πίνακα S είναι οι ιδιοτιμές του πίνακα

A. Τέλος ο πίνακας V είναι ένας ορθογώνιος πίνακας μεγέθους D x r , και περιέχει τα ιδιοδιανύσματα του A^TA.

Με βάση την ανάλυση SVD μπορούμε να αναπαραστήσουμε τον πίνακα ομοιότητας κειμένων ως $A^TA = VS^2V^T$ και τον πίνακα ομοιότητας όρων ως $AA^T = US^2U^T$. Όπως μπορούμε να διαπιστώσουμε, η ομοιότητα και στις δύο περιπτώσεις μπορεί να υπολογιστεί στον r-διάστατο χώρο, όπου r η τάξη του πίνακα A, και όχι στον N-διάστατο ή στον D-διάστατο χώρο. Η διαπίστωση αυτή προκύπτει, αφού μπορούμε να θεωρήσουμε ότι η διανυσματική αναπαράσταση των κειμένων προκύπτει από τις γραμμές του πίνακα VS (μεγέθους D x r), και η διανυσματική αναπαράσταση των όρων προκύπτει από τις γραμμές του πίνακα US (μεγέθους N x r).

Στο LSA μπορεί να γίνει ένα επιπλέον βήμα, επιλέγοντας τα k μεγαλύτερα διαγώνια στοιχεία του πίνακα S (δηλαδή τις k μεγαλύτερες ιδιοτιμές του πίνακα A). Η μειωμένης διάστασης διάσπαση ιδιοτιμών (Reduced SVD) περιγράφεται ως εξής:

$$A_k \;=\; U_k \, S_k \, V_k^T \tag{4.2}$$

, όπου ο πίνακας U_k είναι μεγέθους N x k και περιέχει τις k πρώτες στήλες του πίνακα U, ο πίνακας S_k είναι μεγέθους k x k και περιέχει τα k μεγαλύτερα στοιχεία τις διαγωνίου του S, και τέλος ο πίνακας V_k είναι μεγέθους D x k και περιέχει τις k πρώτες στήλες του πίνακα V. Η παράμετρος k πρέπει να επιλεγεί προσεκτικά, ώστε η προσέγγιση A_k του πίνακα A να μην εισαγάγει πολύ σφάλμα και να είναι δυνατή η ανακατασκευή του πίνακα A χωρίς λάθη. Αν επιλεγεί πολύ μικρή τιμή για το k, υπάρχει απώλεια δεδομένων. Αν επιλεγεί πολύ μεγάλη τιμή για το k, συμπεριλαμβάνουμε περισσότερες από τις απαραίτητες διαστάσεις με αποτέλεσμα την μη αποδοτική μείωση του διανυσματικού χώρου.

Έστω ένα διάνυσμα μίας ερώτησης $\mathbf{q}$, μεγέθους N x 1. Οι συντεταγμένες της ερώτησης στον μειωμένο διανυσματικό χώρο του LSA δίνονται από τον ακόλουθο μετασχηματισμό:

$$\mathbf{q} \;=\; \mathbf{q}^T \, U_k \, S_k^{-1} \tag{4.3}$$

Από τον παραπάνω μετασχηματισμό προκύπτει ένα διάνυσμα μεγέθους 1 x k, το οποίο περιέχει τις συντεταγμένες της ερώτησης **q** στον k-διάστατο μειωμένο διανυσματικό χώρο. Οι συντεταγμένες των κειμένων στον μειωμένο διανυσματικό χώρο βρίσκονται στις γραμμές του πίνακα V_k. Κάθε γραμμή του πίνακα V_k είναι ένα διάνυσμα μεγέθους 1 x k, που περιέχει τις συντεταγμένες του αντίστοιχου κειμένου. Έχοντας τα διανύσματα των κειμένων και το διάνυσμα της ερώτησης, μπορεί να βρεθεί με πιο κείμενο συνδέεται περισσότερο το διάνυσμα της ερώτησης. Η μετρική που χρησιμοποιείται είναι η ομοιότητα συνημιτόνου (cosine similarity). Ας δείξουμε όμως τα παραπάνω με ένα απλό παράδειγμα:

Έστω μια συλλογή που αποτελείται από τα ακόλουθα κείμενα:
d1: Shipment of gold damaged in a fire.
d2: Delivery of silver arrived in a silver truck.
d3: Shipment of gold arrived in a truck.

Στα παραπάνω κείμενα δεν γίνεται αφαίρεση των stop words, όμως αφαιρούνται τα σημεία στίξης και γίνονται όλοι οι χαρακτήρες μικροί. Οι λέξεις παρουσιάζονται με αλφαβητική σειρά στον πίνακα όρων-κειμένων, και δεν γίνεται stemming.

Το ζητούμενο είναι να βαθμολογηθούν τα κείμενα, χρησιμοποιώντας LSA, αναφορικά με την ερώτηση "gold silver truck". Παρακάτω δείχνουμε τον πίνακα όρων-κειμένων καθώς και το διάνυσμα της ερώτησης. Σαν βάρη χρησιμοποιούνται απλά οι αριθμοί των εμφανίσεων των λέξεων στα κείμενα.

Terms		d1	d2	d3		q
a		1	1	1		0
arrived		0	1	1		0
damaged		1	0	0		0
delivery		0	1	0		0
fire		1	0	0		0
gold	A =	1	0	1	q =	1
in		1	1	1		0
of		1	1	1		0
shipment		1	0	1		0
silver		0	2	0		1
truck		0	1	1		1

Έπειτα γίνεται διάσπαση ιδιοτιμών (SVD) στον πίνακα όρων-κειμένων:

$$A = USV^T$$

, όπου:

$$\mathbf{U} = \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix} \quad \mathbf{V^T} = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

Κρατώντας μόνο τις δύο πρώτες στήλες των πινάκων U, V, και τις δύο πρώτες γραμμές και στήλες του πίνακα S, μειώνουμε τις διαστάσεις του προβλήματος σε δύο (Rank 2 Approximation):

$$\mathbf{U} \approx \mathbf{U_k} = \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \quad \mathbf{k = 2} \quad \mathbf{S} \approx \mathbf{S_k} = \begin{bmatrix} 4.0989 & 0.0000 \\ 0.0000 & 2.3616 \end{bmatrix}$$

$$\mathbf{V} \approx \mathbf{V_k} = \begin{bmatrix} -0.4945 & 0.6492 \\ -0.6458 & -0.7194 \\ -0.5817 & 0.2469 \end{bmatrix} \quad \mathbf{V^T} \approx \mathbf{V_k^T} = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \end{bmatrix}$$

Οι συντεταγμένες των κειμένων στον μειωμένο διανυσματικό χώρο, βρίσκονται στις γραμμές του V:

d1(-0.4945, 0.6492)
d2(-0.6458, -0.7194)
d3(-0.5817, 0.2469)

Οι συντεταγμένες της ερώτησης **q** στον μειωμένο διανυσματικό χώρο, βρίσκονται από τον ακόλουθο μετασχηματισμό:

$$\mathbf{q} = \mathbf{q}^T U_k S_k^{-1}$$

$$\mathbf{k} = 2$$

$$\mathbf{q} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \begin{bmatrix} \frac{1}{4.0989} & 0.0000 \\ 0.0000 & \frac{1}{2.3616} \end{bmatrix}$$

$$\mathbf{q} = \begin{bmatrix} -0.2140 & -0.1821 \end{bmatrix}$$

Υπολογίζονται οι ομοιότητες των διανυσμάτων κειμένων με το διάνυσμα της ερώτησης, στον μειωμένο διανυσματικό χώρο:

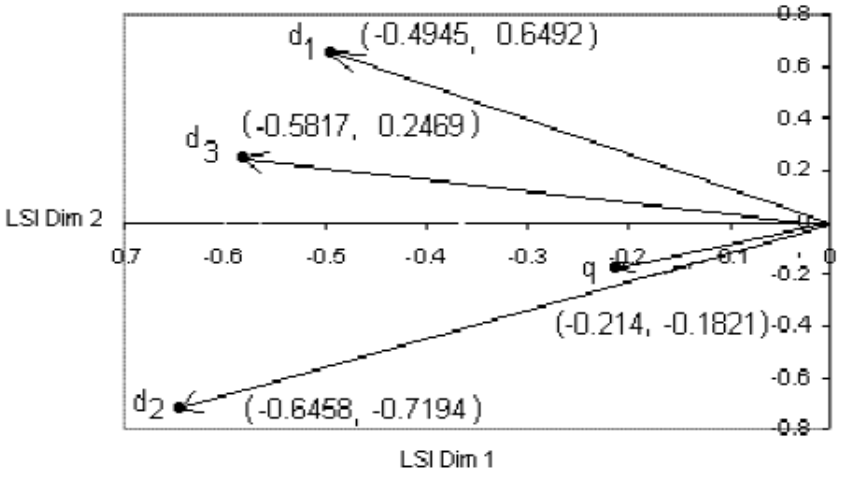
$$\mathbf{sim(q,\ d)} = \frac{\mathbf{q} \bullet \mathbf{d}}{|\,\mathbf{q}\,|\,|\,\mathbf{d}\,|}$$

$$\mathbf{sim(q,\ d_1)} = \frac{(-0.2140)\,(-0.4945) \;+\; (-0.1821)\,(0.6492)}{\sqrt{(-0.2140)^2 + (-0.1821)^2}\ \sqrt{(-0.4945)^2 + (0.6492)^2}} = -0.0541$$

$$\mathbf{sim(q,\ d_2)} = \frac{(-0.2140)\,(-0.6458) \;+\; (-0.1821)\,(-0.7194)}{\sqrt{(-0.2140)^2 + (-0.1821)^2}\ \sqrt{(-0.6458)^2 + (-0.7194)^2}} = 0.9910$$

$$\mathbf{sim(q,\ d_3)} = \frac{(-0.2140)\,(-0.5817) \;+\; (-0.1821)\,(0.2469)}{\sqrt{(-0.2140)^2 + (-0.1821)^2}\ \sqrt{(-0.5817)^2 + (0.2469)^2}} = 0.4478$$

Έτσι κατά σειρά ομοιότητας, με την ερώτηση, τα κείμενα παρουσιάζονται ως εξής: $d2 > d3 > d1$



Όπως φαίνεται και από το παραπάνω γράφημα, το διάνυσμα του d2 βρίσκεται πιο κοντά στο διάνυσμα της ερώτησης σε σχέση με τα διανύσματα των άλλων κειμένων.

Αναφορές

Ο αναγνώστης μπορεί να ανατρέξει στις πηγές [11], [15] για επιπλέον πληροφορίες σχετικά με την λανθάνουσα σημασιολογική ανάλυση.

Κεφάλαιο 5

Σχετική έρευνα

Η δημοσιευμένη εργασία των A. Purandare και D. Litman με τίτλο "Humor: Prosody Analysis and Automatic Recognition for Friends" [2] , ήταν η κινητήριος ιδέα για την έναρξη της εργασίας μας. Στην παρούσα ενότητα θα παρουσιαστεί αναλυτικά αυτή η εργασία, ώστε να υπάρχει ένα μέτρο σύγκρισης με την δική μας εργασία η οποία παρουσιάζεται στα κεφάλαια 6 και 7.

Περίληψη

Στην εργασία αυτή αναλύονται συζητήσεις ομιλίας από την κωμική σειρά Friends. Η ανάλυση γίνεται με την εξαγωγή ηχητικών-προσωδιακών και γλωσσικών χαρακτηριστικών και την εξέταση της χρησιμότητας τους στην αυτόματη αναγνώριση του χιούμορ. Χρησιμοποιείται ένα απλό σχήμα σχολιασμού (annotation scheme), στο οποίο οι σειρές ομιλητών (speaker turns) οι οποίες ακολουθούνται από μηχανικό γέλιο (artificial laugh) θεωρούνται ως χιουμοριστικές και οι υπόλοιπες ως μη χιουμοριστικές. Η εργασία αποκαλύπτει τις σημαντικές διαφορές που υπάρχουν στα προσωδιακά χαρακτηριστικά (όπως pitch, energy, tempo, κ.τ.λ) ανάμεσα σε χιουμοριστική και μη χιουμοριστική ομιλία. Για την αναγνώριση του χιούμορ χρησιμοποιούνται τεχνικές εκμάθησης με επίβλεψη.

Τα δεδομένα και ο σχολιασμός τους

Για την δημιουργία του corpus επιλέγονται διάλογοι από την κωμική σειρά Friends. Συγκεκριμένα επιλέγονται 75 διάλογοι (σκηνές) από έξι επεισόδια της

σειράς Friends (τέσσερα από την πρώτη σεζόν και δύο από την δεύτερη σεζόν). Έτσι συγκεντρώνονται περίπου 2 ώρες audio. Κάθε αρχείο ήχου χωρίζεται (χειροκίνητα), επισημαίνοντας τα όρια των speaker turns με την βοήθεια του εργαλείου wavesurfer. Χρησιμοποιείται ένα απλό σχήμα σχολιασμού (annotation scheme) στο οποίο τα speaker turns τα οποία ακολουθούνται από μηχανικό γέλιο παίρνουν την ετικέτα humorous και τα υπόλοιπα την ετικέτα non-humorous. Από την ανάλυση εξαιρούνται τα διαστήματα μηχανικού γέλιου, διαστήματα σιωπής μεγαλύτερα του ενός δευτερολέπτου, και διαστήματα που περιέχουν μη λεκτικούς ήχους (όπως ήχοι από κουδούνια, μουσική κ.τ.λ). Εν συντομία οι μη λεκτικοί ήχοι που ακολουθούνται από μηχανικό γέλιο δεν θεωρούνται ως χιουμοριστικοί. Με αυτόν τον τρόπο εξαλείφονται καταστάσεις στις οποίες το χιούμορ εκφράζεται μόνο από οπτικά στοιχεία όπως χειρονομίες και εκφράσεις προσώπου. Να σημειώσουμε ότι δεν χρησιμοποιούνται ειδικά φίλτρα για την απομάκρυνση του μη λεκτικού ήχου που μπορεί να παρεμβάλλεται με τα speaker turns. Παρόλα αυτά εάν το μηχανικό γέλιο παρεμβάλλεται με το speaker turn, τότε το speaker turn κόβεται έτσι ώστε να μην συμπεριλαμβάνονται διαστήματα μηχανικού γέλιου. Με άλλα λόγια τα speaker turns καθαρίζονται από τις τυχόν παρεμβολές με τα διαστήματα μηχανικού γέλιου ώστε η μετέπειτα προσωδιακή ανάλυση να είναι δίκαιη. Από την παραπάνω διαδικασία παράγονται συνολικά 1629 speaker turns από τα οποία 714(43.8%) είναι χιουμοριστικά και 915(56.2%) είναι μη χιουμοριστικά. Τέλος έχει ελεγχθεί ότι υπάρχει ένα προς ένα αντιστοιχία μεταξύ των υποτίτλων των speaker turns και των τμημάτων ήχου (audio segments).

Κατανομή ομιλητών

Υπάρχουν 6 βασικοί ηθοποιοί/ομιλητές (3 άνδρες και 3 γυναίκες) καθώς και ένας αριθμός (26) guest ηθοποιών οι οποίοι ομαδοποιούνται, λόγο του πλήθους και της μικρής ατομικής συνεισφοράς τους, σε μία κοινή κλάση GUEST. Επίσης οι περιπτώσεις των speaker turns, στις οποίες πολλαπλοί ηθοποιοί μιλάνε συγχρόνως, ομαδοποιούνται σε μία κλάση MULTI. Τα στατιστικά στοιχεία τα οποία δείχνουν την συνεισφορά κάθε κλάσης στον συνολικό αριθμό των turns, όπως επίσης και την συνεισφορά κάθε κλάσης στα χιουμοριστικά turns, φαίνονται στον πίνακα 5.1.

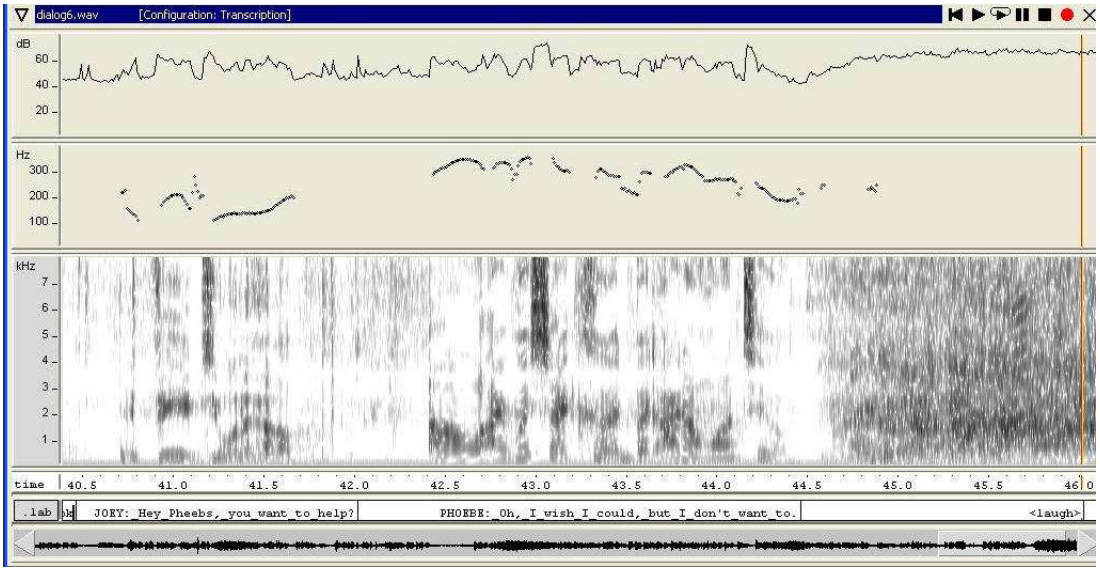
Όπως βλέπουμε από τον πίνακα, ο αριθμός των turns μοιράζεται δίκαια στους 6 κύριους ηθοποιούς. Επίσης παρόλο που η ατομική συνεισφορά κάθε guest ηθοποιού είναι μικρότερη από 5% στα δεδομένα, η συνολική συνεισφορά τους είναι αρκετά μεγάλη (16% των συνολικών turns). Στον πίνακα 5.2 ομαδοποιούνται οι 6 κύριοι ηθοποιοί σε κλάσεις male/female και φαίνεται ότι σχηματίζουν το 83% των συνολικών δεδομένων. Παρατηρείται ότι από τα συνολικά 714 χιουμοριστικά turns, τα 615 (86%) αφορούν τους κύριους ηθοποιούς. Επίσης η κατανομή των turns σε σχέση με το φύλο είναι δίκαιη, με το 50.5% να προκύπτει από τους άνδρες και το 49.5% από τις γυναίκες. Επιπλέον βλέπουμε ότι το 50.6% των male turns ανήκουν στην κατηγορία των χιουμοριστικών. Το ποσοστό των χιουμοριστικών female turns είναι 39.9%. Να σημειωθεί ότι στην ανάλυση που έγινε στον πίνακα 5.2 δεν έλαβαν μέρος οι κλάσεις GUEST, MULTI.

Speaker	# Turns(%)	# Humor(%)
Chandler (M)	244(15)	163(22.8)
Joey (M)	153(9.4)	57(8)
Monica (F)	219(13.4)	74(10.4)
Phoebe (F)	180(11.1)	104(14.6)
Rachel (F)	273(16.8)	90(12.6)
Ross (M)	288(17.7)	127(17.8)
GUEST (26)	263(16.1)	95(13.3)
MULTI	9(0.6)	4(0.6)

Πίνακας 5.1: Speaker Distribution

Speaker	# Turns	# Humor
Male	685 (50.5% of Main)	347 (50.6% of Male)
Female	672 (49.5% of Main)	268 (39.9% of Female)
Total Main	1357 (83.3% of Total)	615 (86.1% of Humor)

Πίνακας 5.2: Gender Distribution for Main Actors



Σχήμα 5.1: Το εργαλείο Wavesurfer

Εξαγωγή χαρακτηριστικών

Από την έρευνα στην ανάλυση συναισθημάτων φωνής έχει δειχθεί ότι προσωδιακά χαρακτηριστικά όπως pitch, energy, speaking rate (tempo) είναι χρήσιμοι δείκτες συναισθηματικών καταστάσεων όπως οργή, ευχαρίστηση, φόβος, πλήξη, κ.τ.λ. Παρότι το χιούμορ δεν θεωρείται απαραίτητα σαν συναισθηματική κατάσταση, παρατηρείται ότι ο χιουμοριστικός λόγος συχνά παρουσιάζει στοιχεία παρόμοια με αυτά του συναισθηματικού λόγου. Στην εργασία χρησιμοποιούνται ηχητικά - προσωδιακά χαρακτηριστικά, καθώς και μη ηχητικά-προσωδιακά χαρακτηριστικά:

Acoustic-Prosodic Features:

- Pitch (F0): Mean, Max, Min, Range, Standard Deviation
- Energy (RMS): Mean, Max, Min, Range, Standard Deviation
- Temporal: Duration, Internal Silence, Tempo

Non Acoustic-Prosodic Features:

- Lexical

- Turn Length (# Words)
- Speaker

Τα ηχητικά-προσωδιακά χαρακτηριστικά του σήματος της φωνής υπολογίζονται με την βοήθεια του Wavesurfer. Στο σχήμα 5.1 φαίνεται το εργαλείο Wavesurfer καθώς και τα πλαίσια που αφορούν την ενέργεια (energy(dB)), την οξύτητα (pitch(Hz)), τον χρόνο, και την περιγραφή (.lab). Στο .lab πλαίσιο βλέπουμε την αντιστοιχία με κείμενο που έχει κάθε dialog turn, καθώς και τα όρια κάθε turn (turn boundaries). Όλα τα χαρακτηριστικά υπολογίζονται στο επίπεδο του turn. Συγκεκριμένα μετρώνται η μέση τιμή (mean value), η μέγιστη τιμή, η ελάχιστη τιμή, η διακύμανση (range), και η τυπική απόκλιση (standard deviation) της τιμής του χαρακτηριστικού (F0 ή RMS) σε όλο το turn (αγνοώντας τα μηδενικά). Η διάρκεια (duration) μετράται σε δευτερόλεπτα ξεκινώντας από την αρχή του turn και καταλήγοντας στο τέλος του, συμπεριλαμβάνοντας τυχόν παύσεις στο ενδιάμεσο. Η εσωτερική σιωπή (internal silence) ενός turn μετράται ως το ποσοστό των frames μηδενικής οξύτητας (zero F0 frames) που υπάρχουν στο turn. Ο ρυθμός (tempo) ενός turn υπολογίζεται ως το κλάσμα του συνολικού αριθμού των συλλαβών προς την διάρκεια του turn. Τα λεκτικά χαρακτηριστικά ενός turn είναι απλά όλες οι λέξεις του turn (αλφαριθμητικά, χωρίς την αφαίρεση των stop words και των αποστρόφων). Η τιμή ενός λεκτικού χαρακτηριστικού ενός turn είναι απλά ο αριθμός των εμφανίσεων της λέξης στο turn. Τέλος το μήκος ενός turn (turn length) είναι ίσο με το πλήθος των λέξεων στο turn.

Προσωδιακή ανάλυση του χιούμορ

Στον πίνακα 5.3 παρουσιάζονται οι μέσες τιμές των διάφορων ηχητικών - προσωδιακών χαρακτηριστικών, υπολογισμένες από όλα τα speaker turns κάθε κατηγορίας (Humor, Non-Humor). Τα χαρακτηριστικά τα οποία έχουν στατιστικά σημαντική διαφορά, ανάμεσα στις δύο κατηγορίες, μαρκάρονται στον πίνακα με αστερίσκο. Όπως φαίνεται όλα τα χαρακτηριστικά εκτός από τα Mean-F0, StdDev-F0, παρουσιάζουν σημαντικές διαφορές ανάμεσα στις δύο κατηγορίες. Ο πίνακας 5.3 δείχνει ότι τα χιουμοριστικά turns είναι μακρύτερα και σε διάρκεια και σε

πλήθος λέξεων. Επίσης τα χιουμοριστικά turns έχουν μικρότερο internal-silence και υψηλότερο ρυθμό (tempo). Τα χαρακτηριστικά που αφορούν την οξύτητα (Pitch(F0)) και την ενέργεια (energy(RMS)) έχουν μεγαλύτερες μέγιστες αλλά μικρότερες ελάχιστες τιμές για την κατηγορία Humor. Η προηγούμενη παρατήρηση εξηγεί τις μεγαλύτερες (σε σχέση με την κατηγορία Non-Humor) τιμές διακύμανσης και τυπικής απόκλισης που παρουσιάζονται στην κατηγορία Humor.

Feature	Humor	Non-Humor
Mean-F0	206.9	208.9
Max-F0*	299.8	293.5
Min-F0*	121.1	128.6
Range-F0*	178.7	164.9
StdDev-F0	41.5	41.1
Mean-RMS*	58.3	57.2
Max-RMS*	76.4	75
Min-RMS*	44.2	44.6
Range-RMS*	32.16	30.4
StdDev-RMS*	7.8	7.5
Duration*	3.18	2.66
Int-Sil*	0.452	0.503
Tempo*	3.21	3.03
Length*	10.28	7.97

Πίνακας 5.3: Humor Prosody: Mean feature values for Humor and Non-Humor groups

Η επίδραση του φύλου στην προσωδιακή ανάλυση του χιούμορ

Για την ανάλυση της προσωδίας του χιούμορ ανάμεσα στα δύο φύλα, διεξάγεται ένα ένα 2-way ANOVA test. Το φύλο (male/female) και το humor (yes/no) χρησιμοποιούνται σαν αμετάβλητες παράμετροι και καθένα από τα προσωδιακά χαρακτηριστικά σαν εξαρτημένη μεταβλητή. Το test φανερώνει την επίδραση (στην προσωδία) του χιούμορ προσαρμοσμένη για το φύλο, την επίδραση του φύλου (στην προσωδία) προσαρμοσμένη για το χιούμορ και τέλος την αλληλεπίδραση του φύλου και του χιούμορ στην προσωδία (αν η επίδραση του χιούμορ στην προσωδία διαφέρει ανάλογα με το φύλο). Στον πίνακα 5.4 παρουσιάζονται τα αποτελέσματα του 2-way

ANOVA test, όπου το Y σημαίνει σημαντική επίδραση και το N ασήμαντη επίδραση.

Feature	Humor	Gender	Humor x Gender
Mean-F0	N	Y	N
Max-F0	Y	Y	Y
Min-F0	Y	Y	Y
Range-F0	Y	Y	N
StdDev-F0	N	Y	Y
Mean-RMS	Y	Y	N
Max-RMS	Y	Y	N
Min-RMS	Y	Y	N
Range-RMS	Y	Y	N
StdDev-RMS	Y	Y	N
Duration	Y	Y	N
Int-Sil	Y	N	N
Tempo	Y	N	N
Length	Y	Y	N

Πίνακας 5.4: Gender Effect on Humor Prosody: 2-way ANOVA Results

Αναλύοντας τον παραπάνω πίνακα βλέπουμε, για παράδειγμα, ότι το tempo διαφέρει σημαντικά ανάμεσα στις κατηγορίες humor και non-humor, αλλά όχι ανάμεσα στα δύο φύλα, επίσης φαίνεται ότι δεν υπάρχει αλληλεπίδραση του χιούμορ και του φύλου στο tempo. Όπως προηγουμένως, όλα τα χαρακτηριστικά εκτός από τα Mean-F0, StdDev-F0, παρουσιάζουν σημαντικές διαφορές ανάμεσα στα humor και non-humor groups. Επίσης όλα τα χαρακτηριστικά εκτός από τα internal silence, tempo, παρουσιάζουν σημαντικές διαφορές ανάμεσα στα δύο φύλα. Ακόμη βλέπουμε ότι μόνο τα χαρακτηριστικά Max-F0, Min-F0, StdDev-F0, δείχνουν ότι υπάρχει σημαντική αλληλεπίδραση του φύλου με το χιούμορ. Με άλλα λόγια η επίδραση του χιούμορ στα χαρακτηριστικά αυτά είναι εξαρτημένη από το φύλο. Για να αποδειχθεί αυτό, υπολογίζονται οι μέσες τιμές των διάφορων χαρακτηριστικών ξεχωριστά για τις κατηγορίες male/female (τα αποτελέσματα φαίνονται στους πίνακες 5.5, 5.6).

Feature	Humor	Non-Humor
Mean-F0*	188.14	176.43
Max-F0*	276.94	251.7
Min-F0	114.54	113.56
Range-F0*	162.4	138.14
StdDev-F0*	37.83	34.27
Mean-RMS*	57.86	56.4
Max-RMS*	75.5	74.21
Min-RMS	44.04	44.12
Range-RMS*	31.46	30.09
StdDev-RMS*	7.64	7.31
Duration*	3.1	2.57
Int-Sil*	0.44	0.5
Tempo*	3.33	3.1
Length*	10.27	8.1

Πίνακας 5.5: Humor Prosody for Male Speakers

Feature	Humor	Non-Humor
Mean-F0	235.79	238.75
Max-F0*	336.15	331.14
Min-F0*	133.63	143.14
Range-F0*	202.5	188
StdDev-F0	46.33	46.6
Mean-RMS*	58.44	57.64
Max-RMS*	77.33	75.57
Min-RMS*	44.08	44.74
Range-RMS*	33.24	30.83
StdDev-RMS*	8.18	7.59
Duration*	3.35	2.8
Int-Sil*	0.47	0.51
Tempo	3.1	3.1
Length*	10.66	8.25

Πίνακας 5.6: Humor Prosody for Female Speakers

Από τους πίνακες 5.5, 5.6 φαίνεται ότι οι άνδρες ομιλητές έχουν μεγαλύτερες τιμές στα pitch features (Mean-F0, Min-F0, StdDev-F0), όταν εκφράζουν το χιούμορ, ενώ οι γυναίκες έχουν μικρότερες. Για τους άνδρες ομιλητές οι διαφορές των Min-F0, Min-RMS στα humor και non-humor groups δεν είναι στατιστικά σημαντικές, ενώ για τις γυναίκες, τα χαρακτηριστικά Mean-F0, StdDev-F0, tempo δεν παρουσιάζουν στατιστικά σημαντικές διαφορές στα humor/non-humor groups. Παρατηρείται ότι οι διαφορές (που παρουσιάζουν οι άνδρες) στα Mean-F0, Max-F0, Range-F0, ανάμεσα στα humor/non-humor groups είναι αρκετά μεγαλύτερες σε σχέση με αυτές που παρουσιάζουν οι γυναίκες. Κλείνοντας σημειώνεται ότι παρόλο που τα ηχητικά-προσωδιακά χαρακτηριστικά διαφέρουν μεταξύ ανδρών και γυναικών, το προσωδιακό στυλ έκφρασης του χιούμορ διαφέρει μόνο σε μερικά pitch-features (και στο μέγεθος και στην κατεύθυνση).

Η επίδραση των ομιλητών στην προσωδιακή ανάλυση του χιούμορ

Για την ανάλυση της επίδρασης των ομιλητών στην προσωδία του χιούμορ πραγματοποιείται ένα παρόμοιο ANOVA test με το προηγούμενο. Το humor (yes/no) και ο ομιλητής (8 groups όπως φαίνεται στον πίνακα 5.1) θεωρούνται ως αμετάβλητες παράμετροι και κάθε ηχητικό-προσωδιακό χαρακτηριστικό θεωρείται σαν μία εξαρτημένη μεταβλητή για ένα 2-way ANOVA test. Στον πίνακα 5.7 παρουσιάζονται τα αποτελέσματα της ανάλυσης. Ο πίνακας δείχνει την επίδραση του χιούμορ προσαρμοσμένη για τον ομιλητή, την επίδραση του ομιλητή προσαρμοσμένη για το χιούμορ, και την αλληλεπίδραση χιούμορ και ομιλητή, σε καθένα ηχητικό-προσωδιακό χαρακτηριστικό. Σύμφωνα με τον πίνακα 5.7 δεν υπάρχει πλέον επίδραση του χιούμορ στα Min-F0, Mean-RMS, Tempo, υπό την παρουσία της μεταβλητής του ομιλητή. Από την άλλη πλευρά ο ομιλητής έχει σημαντική επίδραση σε όλα τα προσωδιακά χαρακτηριστικά. Τέλος παρατηρείται ότι αλληλεπίδραση χιούμορ και ομιλητή έχει σημαντική επίδραση μόνο στα pitch features Mean-F0, Max-F0, Min-F0, δηλαδή η επίδραση του χιούμορ σε αυτά χαρακτηριστικά διαφέρει από ομιλητή σε ομιλητή.

Feature	Humor	Speaker	Humor x Speaker
Mean-F0	N	Y	Y
Max-F0	Y	Y	Y
Min-F0	N	Y	Y
Range-F0	Y	Y	N
StdDev-F0	N	Y	N
Mean-RMS	N	Y	N
Max-RMS	Y	Y	N
Min-RMS	Y	Y	N
Range-RMS	Y	Y	N
StdDev-RMS	Y	Y	N
Duration	Y	Y	N
Int-Sil	Y	Y	N
Tempo	N	Y	N
Length	Y	Y	N

Πίνακας 5.7: Speaker Effect on Humor Prosody: 2-way ANOVA Results

Αναγνώριση του χιούμορ χρησιμοποιώντας τεχνικές εκμάθησης με επίβλεψη

Χρησιμοποιούνται τυπικοί machine learning classifiers για την αυτόματη ταξινόμηση των speaker turns στις κατηγορίες humor/non-humor. Χρησιμοποιείται ο αλγόριθμος decision tree (ADTree from Weka) και εκτελείται ένα 10-fold cross validation πείραμα πάνω σε όλα τα 1629 turns των δεδομένων. Το baseline για αυτά τα πειράματα είναι το 56.2% (το ποσοστό της πολυπληθέστερης κατηγορίας (non-humorous)). Στον πίνακα 5.8 παρουσιάζονται τα αποτελέσματα της ταξινόμησης για έξι κατηγορίες χαρακτηριστικών: lexical alone, lexical+speaker, prosody alone, prosody+speaker, lexical+prosody, lexical+prosody+speaker (all). Συνολικά υπάρχουν 2025 χαρακτηριστικά, από τα οποία τα 2011 είναι λεκτικά (τύποι λέξεων συν το turn length), τα 13 είναι ηχητικά-προσωδιακά, και το 1 η πληροφορία για τον ομιλητή. Όλα τα αποτελέσματα, όπως φαίνεται και από τον πίνακα 5.8, βρίσκονται πάνω από το baseline. Παρατηρείται ότι η ακρίβεια του ταξινομητή βελτιώνεται προσθέτοντας την πληροφορία του ομιλητή (και στα λεκτικά και στα

προσωδιακά χαρακτηριστικά). Είναι επίσης ενδιαφέρων ότι μόλις 13 προσωδιακά χαρακτηριστικά προσφέρουν συγκρίσιμα αποτελέσματα με αυτά που προέρχονται από 2011 λεκτικά χαρακτηριστικά. Το σχήμα 5.2 δείχνει το decision tree που παράγεται από τον ταξινομητή στις 10 πρώτες επαναλήψεις. Οι αριθμοί δείχνουν την σειρά με την οποία δημιουργούνται οι κόμβοι και οι οδοντώσεις τις σχέσεις parent-child. Παρατηρείται ότι στις πρώτες 10 επαναλήψεις επιλέγονται τα χαρακτηριστικά του ομιλητή και τα προσωδιακά χαρακτηριστικά, ενώ τα λεκτικά χαρακτηριστικά επιλέγονται μετέπειτα (δεν φαίνεται στο σχήμα). Το γεγονός αυτό είναι σύμφωνο με την υπόθεση ότι φωνητικά χαρακτηριστικά είναι καλύτερα, σε σχέση με τα λεκτικά χαρακτηριστικά, στον διαχωρισμό των turns σε κατηγορίες humorous/non-humorous. Στο σχήμα 5.3 φαίνεται πως επηρεάζεται η ακρίβεια του ταξινομητή σχετικά με τον όγκο των δεδομένων. Παρατηρείται ότι η ακρίβεια του ταξινομητή δεν είναι ευαίσθητη στην ποσότητα των δεδομένων. Στον πίνακα 5.9 φαίνονται τα αποτελέσματα της ταξινόμησης (με την χρησιμοποίηση όλων των features) σε σχέση με το φύλο.

Feature	-Speaker	+Speaker
Lex	61.14 (2011)	63.5 (2012)
Prosody	60 (13)	63.8 (14)
Lex + Prosody	62.6 (2024)	64 (2025)

Πίνακας 5.8: Humor Recognition Results (% Correct)

Gender	Baseline	Classifier
Male	50.6	64.63
Female	60.1	64.8

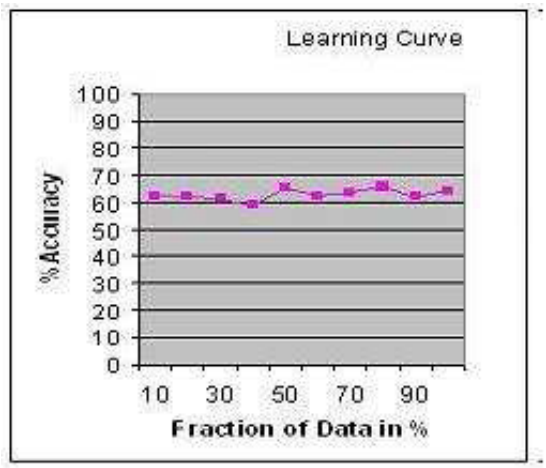
Πίνακας 5.9: Humor Recognition Results by Gender

Αναφορές

Ο αναγνώστης μπορεί να ανατρέξει στις πηγές [9], [4], [6], [8], [5], [7], [3], για επιπλέον πληροφορίες σχετικά με την έρευνα στην αναγνώριση του χιούμορ.

```
|   (1)SPEAKER = chandler: 0.469
|   (1)SPEAKER != chandler: -0.083
|   |   (4)SPEAKER = phoebe: 0.373
|   |   (4)SPEAKER != phoebe: -0.064
|   (2)DURATION < 1.515: -0.262
|   |   (5)SILENCE < 0.659: 0.115
|   |   (5)SILENCE >= 0.659: -0.465
|   |   (8)SD_F0 < 9.919: -1.11
|   |   (8)SD_F0 >= 9.919: 0.039
|   (2)DURATION >= 1.515: 0.1
|   |   (3)MEAN_RMS < 56.117: -0.274
|   |   (3)MEAN_RMS >= 56.117: 0.147
|   |   |   (7)come < 0.5: -0.056
|   |   |   (7)come >= 0.5: 0.417
|   |   (6)SD_F0 < 57.333: 0.076
|   |   (6)SD_F0 >= 57.333: -0.285
|   |   (9)MAX_RMS < 86.186: 0.011
|   |   |   (10)MIN_F0 < 166.293: 0.047
|   |   |   (10)MIN_F0 >= 166.293: -0.351
|   |   (9)MAX_RMS >= 86.186: -0.972
Legend: +ve = humor, -ve = non-humor
```

Σχήμα 5.2: Δέντρο απόφασης (φαίνονται μόνο οι 10 πρώτες επαναλήψεις)



Σχήμα 5.3: Καμπύλη εκμάθησης: % Accuracy versus % Fraction of Data

Μέρος II

Η ΕΡΓΑΣΙΑ

Κεφάλαιο 6

Αυτόματη εξαγωγή χιουμοριστικών και μη χιουμοριστικών υποτίτλων

Σε αυτό το κεφάλαιο θα περιγράψουμε τον τρόπο με τον οποίο υλοποιήσαμε ένα σύστημα το οποίο διαχωρίζει αυτόματα τους χιουμοριστικούς, από τους μη χιουμοριστικούς υπότιτλους, σε σειρές στις οποίες χρησιμοποιείται μηχανικό γέλιο (artificial laugh). Στην ενότητα 6.1 περιγράφουμε δεδομένα μας, καθώς και την διαδικασία επισήμανσης των τμημάτων ήχου (audio segments) μηχανικού γέλιου. Στην ενότητα 6.2 περιγράφουμε την εκπαίδευση του μοντέλου μας (μείξη κανονικών κατανομών) και την ταξινόμηση των δεδομένων μας (πλαίσια ήχου (audio frames)) σε κατηγορίες μηχανικού γέλιου/μη μηχανικού γέλιου. Τέλος στην ενότητα 6.3 περιγράφεται ο διαχωρισμός των υποτίτλων σε κατηγορίες χιουμοριστικών/μη χιουμοριστικών υποτίτλων.

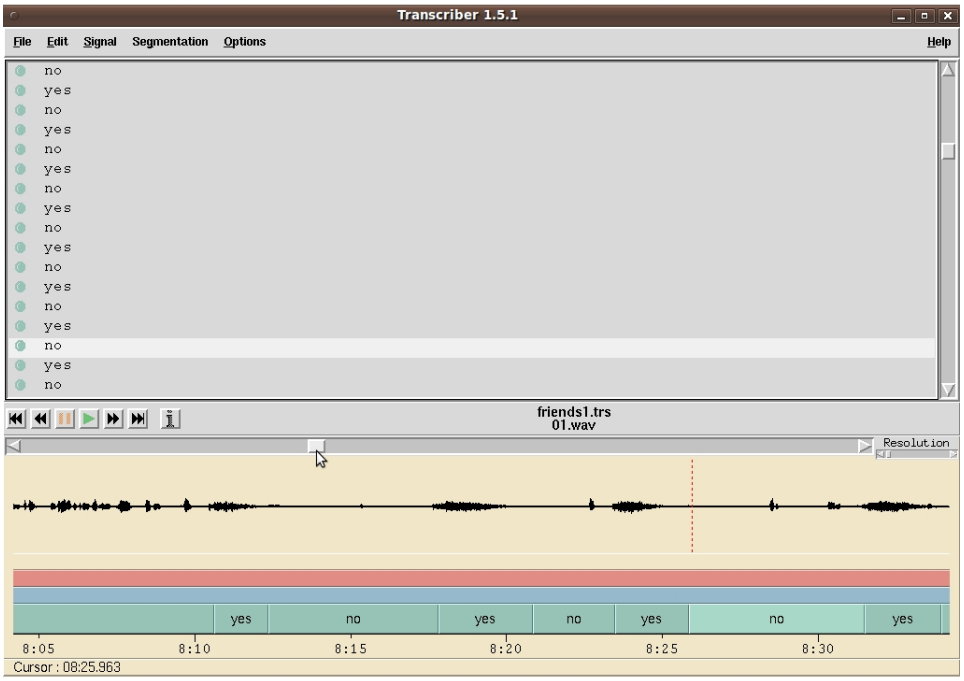
6.1 Τα δεδομένα μας

Τα δεδομένα μας είναι 23 avi και 23 srt αρχεία, τα οποία αντιστοιχούν στα video και τους υπότιτλους από τα 23 επεισόδια της 4ης σεζόν της δημοφιλούς σειράς Friends. Από τα 23 avi αρχεία, γίνεται εξαγωγή του ήχου (audio) με την βοήθεια

του mplayer. Έτσι παράγονται 23 wav αρχεία τα οποία έχουν τα εξής χαρακτηριστικά:

- Μέγεθος δείγματος ήχου: 16 bit
- Κανάλια: μονοφωνικά
- Ρυθμός δειγμάτων ήχου: 8KHz
- Μορφή ήχου: PCM

Στην συνέχεια με την βοήθεια του εργαλείου transcriber εντοπίζουμε τα διαστήματα μηχανικού γέλιου, στα τρία πρώτα επεισόδια. Στην επόμενη εικόνα, φαίνεται το σήμα του ήχου ενός επεισοδίου καθώς και ο χαρακτηρισμός των διαστημάτων χρόνου με ετικέτες (“yes” για τα διαστήματα μηχανικού γέλιου και “no” για τα λοιπά διαστήματα).



Το εργαλείο transcriber παράγει σαν έξοδο trs αρχεία της μορφής:

<?xml version=”1.0” encoding=”ISO-8859-1”?>

<!DOCTYPE Trans SYSTEM ”trans-14.dtd”>

<Trans scribe=" " audio_filename="friends1" version="3"

version_date="090510">

<Episode>

<Section type="report" startTime="0" endTime="1411.180">

<Turn startTime="0" endTime="1411.180">

<Sync time="0"/>

no

<Sync time="11.287"/>

yes

<Sync time="13.323"/>

...

Με κώδικα σε perl "καθαρίζουμε" τα trs αρχεία από την περιττή πληροφορία και δημιουργούμε πίνακες διαστημάτων χρόνου - κατηγοριών της μορφής του πίνακα 6.1. Οι γραμμές κάθε πίνακα περιέχουν τα διαστήματα χρόνου μαζί με τις αντίστοιχες ετικέτες κατηγορίας. Έτσι τελικά δημιουργούμε τρεις πίνακες (για τρία πρώτα επεισόδια), οι οποίοι περιέχουν την πληροφορία της κατηγορίας για κάθε διάστημα χρόνου που ανήκει σε αυτά τα επεισόδια. Με N συμβολίζουμε το πλήθος των διαστημάτων χρόνου.

Number of audio segment	Start time (sec)	End time (sec)	Category (1 για διαστήματα artificial laugh και 0 για τα λοιπά διαστήματα)
1	0	11.287	0
2	11.287	13.323	1
⋮	⋮	⋮	⋮
N	...	...	...

Πίνακας 6.1: Labels of audio segments

6.2 Εκπαίδευση μοντέλου - Ταξινόμηση δεδομένων

6.2.1 Δημιουργία των δεδομένων εκπαίδευσης και ο διαχωρισμός τους σε κατηγορίες μηχανικού γέλιου/μη μηχανικού γέλιου

Το πρώτο βήμα για την εκπαίδευση του μοντέλου μας (μείξη κανονικών κατανομών) είναι η δημιουργία των δεδομένων εκπαίδευσης. Με την βοήθεια του εργαλείου HTK, γίνεται εξαγωγή των Mel-frequency cepstral coefficients (MFCCs) για τα 23 αρχεία ήχου (ενότητα 6.1). Το σήμα του ήχου χωρίζεται σε διαδοχικά πλαίσια (frames), μήκους 0.02 sec (20 msec). Η περίοδος του frame είναι 0.01 sec (10 msec). Αυτό σημαίνει ότι κάθε frame επικαλύπτεται με το επόμενο κατά 0.01sec. Κάθε frame πολλαπλασιάζεται με την συνάρτηση Hamming. Τέλος για κάθε frame παράγονται 39 MFCCs, οι οποίοι δίνουν μία συμπαγή αναπαράσταση των φασματικών ιδιοτήτων του frame:

- Οι 12 πρώτοι MFCCs: [c1, . . . ,c12]
- Ο "null" MFCC: c0, ο οποίος είναι αναλογικός της συνολικής ενέργειας του frame.
- 13 "Delta coefficients", οι οποίοι αποτιμούν την πρώτη παράγωγο των συντελεστών [c0, c1, . . . , c12].
- 13 "Acceleration coefficients", οι οποίοι αποτιμούν την δεύτερη παράγωγο των συντελεστών [c0, c1, . . . , c12].

Το επόμενο βήμα είναι να αναθέσουμε την ετικέτα "1", στα frames που ανήκουν στην κατηγορία μηχανικού γέλιου, και την ετικέτα "0", στα frames που δεν ανήκουν στην κατηγορία μηχανικού γέλιου. Η διαδικασία αυτή θα γίνει για τα frames των τριών πρώτων επεισοδίων, αφού σε αυτά έχει γίνει το χειροκίνητο labeling των τμημάτων μηχανικού γέλιου. Για να προσδιορίσουμε την κατηγορία στην οποία ανήκει κάθε frame πρέπει να περάσουμε από το επίπεδο διαστημάτων χρόνου, στα

οποία έχει γίνει το labeling, σε επίπεδο διαστημάτων frames. Όπως προαναφέραμε στην ενότητα 6.1, έχουμε δημιουργήσει (για τα τρία πρώτα επεισόδια) πίνακες διαστημάτων χρόνου - κατηγοριών της μορφής:

Number of audio segment	Start time (sec)	End time (sec)	Category (1 για διαστήματα artificial laugh και 0 για τα λοιπά διαστήματα)
1	0	11.287	0
2	11.287	13.323	1
⋮	⋮	⋮	⋮
N	...	...	...

Διαιρώντας τους παραπάνω χρόνους με την περίοδο του frame (0.01 sec) και κρατώντας το ακέραιο κομμάτι, προσέχοντας παράλληλα να μην επικαλύπτονται τα διαστήματα των frames που θα πάρουμε, δημιουργούμε πίνακες διαστημάτων frames - κατηγοριών της μορφής του πίνακα 6.2. Με το N συμβολίζουμε το πλήθος των διαστημάτων που χαρακτηρίστηκαν (με τις ετικέτες μηχανικού/μη μηχανικού γέλιου).

Number of audio segment	Number of start frame	Number of end frame	Category (1 για διαστήματα artificial laugh και 0 για τα λοιπά διαστήματα)
1	0	1127	0
2	1128	1331	1
⋮	⋮	⋮	⋮
N	...	...	...

Πίνακας 6.2: Labels διαστημάτων frames

Με τις παραπάνω πληροφορίες για τα frames δημιουργούμε τρεις πίνακες της μορφής του πίνακα 6.3 (που αντιστοιχούν στα τρία πρώτα επεισόδια), ο καθένας από τους οποίους έχει πλήθος γραμμών ίσο με το πλήθος των frames του αντίστοιχου επεισοδίου. Ο αριθμός των στηλών κάθε πίνακα είναι 40. Το πρώτο στοιχείο κάθε γραμμής αντιπροσωπεύει την κατηγορία του frame, και τα υπόλοιπα 39 είναι οι MFCCs του frame. Με K συμβολίζουμε το πλήθος των frames του επεισοδίου.

Number of frame	Category (1 για frames artificial laugh και 0 για τα υπόλοιπα frames)	MFCCs vectors ([c1 .. c39])
0	0	...
1	0	...
2	0	...
⋮	⋮	⋮
K	...	...

Πίνακας 6.3: Labels of mfccs vectors

Τα δεδομένα της κατηγορίας μηχανικού γέλιου, είναι τα διανύσματα των MFCCs των γραμμών που έχουν την ετικέτα (πρώτο στοιχείο της γραμμής) ίση με "1". Τα διανύσματα των MFCCs των υπόλοιπων γραμμών, είναι τα δεδομένα της κατηγορίας μη μηχανικού γέλιου.

Αναφορές

Ο αναγνώστης μπορεί να ανατρέξει στην πηγή [14], όπου υπάρχει μία πλήρης τεκμηρίωση του εργαλείου HTK.

6.2.2 Εκπαίδευση της μείξης κανονικών κατανομών και ταξινόμηση των δεδομένων

Το μοντέλο που θα χρησιμοποιήσουμε είναι η μείξη κανονικών κατανομών (Gaussian Mixture Model). Για την εκπαίδευση του μοντέλου και την ταξινόμηση των testing δεδομένων, θα χρησιμοποιήσουμε το netlab toolkit. Ακολουθούν τα επόμενα βήματα:

- Αρχικοποίηση μοντέλου:
  Ο αριθμός των κανονικών κατανομών του μοντέλου μας είναι εννέα. Οι διαστάσεις των δεδομένων του μοντέλου είναι 39, όσα είναι και τα στοιχεία του διανύσματος των MFCCs καθενός frame. Επίσης ο πίνακας συνδιασπορών κάθε κανονικής κατανομής είναι διαγώνιος με τα διαγώνια στοιχεία ίσα με την μονάδα. Κάθε κατανομή αρχικοποιείται με το ίδιο βάρος, και τα βάρη όλων

των κατανομών αθροίζουν στο ένα. Τα κέντρα (μέσες τιμές των κανονικών κατανομών) του μοντέλου μας αρχικοποιούνται τυχαία.

- k-means clustering:

  Έπειτα από την αρχικοποίηση, χρησιμοποιούμε τα δεδομένα μας για να παραμετροποιήσουμε το μοντέλο μας σύμφωνα με αυτά. Χρησιμοποιείται ο αλγόριθμος k-means, για να προσδιορίσουμε τα κέντρα του μοντέλου. Σύμφωνα με την αρχικοποίηση, το μοντέλο μας έχει 9 κανονικές κατανομές ως συνιστώσες (components), οπότε k=9. Έτσι τα δεδομένα μας χωρίζονται σε 9 ομάδες (clusters). Το βάρος κάθε κανονικής κατανομής υπολογίζεται αναλογικά με το πλήθος των παραδειγμάτων που ανήκουν στην αντίστοιχη ομάδα. Οι πίνακες συνδιασποράς κάθε κατανομής υπολογίζονται με βάση τα δεδομένα της αντίστοιχης ομάδας.

- Expectation-Maximization (EM):

  Η παραπάνω παραμετροποίηση χρησιμοποιείται ως αρχικό σημείο, στην εκπαίδευση του μοντέλου μας με τον αλγόριθμο Expectation-Maximization (EM), ώστε να μεγιστοποιήσουμε την πιθανότητα σωστής εκτίμησης των δεδομένων μας. Ο μέγιστος αριθμός των επαναλήψεων του αλγόριθμου στην εφαρμογή μας είναι το 100, και πραγματοποιείται όταν δεν επιτευχθεί το κατώφλι σύγκλισης του αλγορίθμου.

Εφαρμόζοντας τον αλγόριθμο του Cross validation (leave one out) στα τρία πρώτα επεισόδια παίρνουμε τα training/testing sets, που φαίνονται στον πίνακα 6.4.

Number of experiment	Training set	Testing set
1	MFCCs vectors of frames in episodes 1,2	MFCCs vectors of frames in episode 3
2	MFCCs vectors of frames in episodes 1,3	MFCCs vectors of frames in episode 2
3	MFCCs vectors of frames in episodes 2,3	MFCCs vectors of frames in episode 1

Πίνακας 6.4: Training and testing sets

Για κάθε πείραμα, με βάση το αντίστοιχο training set, εκπαιδεύουμε δύο μοντέλα

μείξης κανονικών κατανομών. Το πρώτο είναι για την κατηγορία μηχανικού γέλιου, και εκπαιδεύεται με τα διανύσματα των MFCCs των frames μηχανικού γέλιου. Το δεύτερο είναι για την κατηγορία μη μηχανικού γέλιου, και εκπαιδεύεται με τα διανύσματα των MFCCs των υπόλοιπων frames.

Το επόμενο βήμα είναι ταξινόμηση των frames, από το testing set, σε κατηγορίες μηχανικού γέλιου/μη μηχανικού γέλιου. Ο υπολογισμός των a-priori πιθανοτήτων για κάθε κατηγορία γίνεται ως εξής:

- Για την κατηγορία μηχανικού γέλιου:

$$P(\omega_1) = \frac{\textbf{Αριθμός των frames μηχανικού γέλιου στο training set}}{\textbf{Αριθμός όλων των frames στο training set}}$$

- Για την κατηγορία μη μηχανικού γέλιου:

$$P(\omega_2) = \frac{\textbf{Αριθμός των frames μη μηχανικού γέλιου στο training set}}{\textbf{Αριθμός όλων των frames στο training set}}$$

Έστω ένα frame x, προς ταξινόμηση. Με βάση το διάνυσμα των MFCCs του frame και τα μοντέλα για το μηχανικό γέλιο/μη μηχανικό γέλιο, υπολογίζεται η πιθανότητα του frame με κάθε κατηγορία ($P(x|\omega_1),\ P(x|\omega_2)$). Το frame ταξινομείται στην κατηγορία η οποία δίνει τη μεγαλύτερη πιθανότητα:

$$\widehat{\omega} = \arg\max_{\omega_i} P(\omega_i|x) = \arg\max_{\omega_i}\{P(\omega_i)\ .\ P(x|\omega_i)\}\ ,\ i = 1, 2$$

Μετά από τη ταξινόμηση όλων των frames του testing set, δημιουργείται ένα διάνυσμα, με μήκος ίσο με το πλήθος των frames του testing set. Το διάνυσμα αυτό περιέχει τις αποφάσεις της ταξινόμησης για κάθε frame. Ουσιαστικά είναι ένα διάνυσμα που περιέχει τις τιμές 0, 1. Το ένα αντιστοιχεί σε frames μηχανικού γέλιου, και το μηδέν στα υπόλοιπα frames. Το διάνυσμα αυτό φιλτράρεται από median filter, με μήκος παραθύρου ίσο με 80. Τα διαστήματα χωρίς μηχανικό γέλιο ακολουθούνται από διαστήματα μηχανικού γέλιου. Τα διαστήματα αυτά έχουν πολύ μεγαλύτερη διάρκεια σε σχέση με την περίοδο του frame, οπότε περιμένουμε αρκετά διαδοχικά frames της ίδιας κατηγορίας στο διάνυσμα. Για παράδειγμα αν υπάρχει η τιμή 1 ανάμεσα σε πολλά μηδενικά, είναι πολύ πιθανό το frame που αντιστοιχεί

στην τιμή 1 να έχει ταξινομηθεί λανθασμένα. Το median filter αντιμετωπίζει τέτοιες περιπτώσεις.

Έχοντας το διάνυσμα των αποφάσεων, και την πληροφορία για την κατηγορία στην οποία ανήκει κάθε frame των testing δεδομένων (από το labeling (ενότητα 6.2.1)), μπορούμε να δούμε ποια frames ταξινομήθηκαν σωστά. Έστω tp ο αριθμός των frames μηχανικού γέλιου που ταξινομήθηκαν σωστά και fn ο αριθμός των frames μηχανικού γέλιου που ταξινομήθηκαν λάθος. Αντίστοιχα συμβολίζουμε με tn (fp) τον αριθμό των frames μη μηχανικού γέλιου που ταξινομήθηκαν σωστά (λάθος).

Για την εκτίμηση των αποτελεσμάτων μας θα χρησιμοποιήσουμε τρεις μετρικές:

1. $Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$
2. $Precision = \frac{tp}{tp + fp}$
3. $Recall = \frac{tp}{tp + fn}$

Στον πίνακα 6.5 παρουσιάζονται τα αποτελέσματα. Εφόσον είδαμε ότι ο ταξινομητής έχει ικανοποιητικά ποσοστά επιτυχίας, μπορούμε να προχωρήσουμε στο επόμενο βήμα της πειραματικής διαδικασίας. Στο βήμα αυτό χρησιμοποιούμε τα διανύσματα των MFCCs των frames των τριών πρώτων επεισοδίων σαν training set, και ταξινομούμε τα frames των υπόλοιπων 20 επεισοδίων. Έτσι έχουμε για κάθε επεισόδιο (από τα 23 επεισόδια), ένα διάνυσμα ετικετών με μήκος ίσο με το πλήθος των frames του επεισοδίου, το οποίο περιέχει για κάθε frame του επεισοδίου την ετικέτα της κατηγορίας στην οποία ανήκει.

<table>
<tr><th>Number of experiment</th><th>Misclassification tables</th></tr>
<tr><td>1</td><td>
<table>
<tr><th>true classes/hypothesized classes</th><th>artificial laugh</th><th>non-artificial laugh</th></tr>
<tr><td>artificial laugh</td><td>30397 (tp)</td><td>7647 (fn)</td></tr>
<tr><td>non-artificial laugh</td><td>2355 (fp)</td><td>93927 (tn)</td></tr>
</table>
precision = 0.9281
recall = 0.799
accuracy = 0.92554
</td></tr>
<tr><td>2</td><td>
<table>
<tr><th>true classes/hypothesized classes</th><th>artificial laugh</th><th>non-artificial laugh</th></tr>
<tr><td>artificial laugh</td><td>38955 (tp)</td><td>6129 (fn)</td></tr>
<tr><td>non-artificial laugh</td><td>3176 (fp)</td><td>96070 (tn)</td></tr>
</table>
precision = 0.92462
recall = 0.86405
accuracy = 0.93553
</td></tr>
<tr><td>3</td><td>
<table>
<tr><th>true classes/hypothesized classes</th><th>artificial laugh</th><th>non-artificial laugh</th></tr>
<tr><td>artificial laugh</td><td>29245 (tp)</td><td>4939 (fn)</td></tr>
<tr><td>non-artificial laugh</td><td>3543 (fp)</td><td>103391 (tn)</td></tr>
</table>
precision = 0.89194
recall = 0.85552
accuracy = 0.93989
</td></tr>
<tr><td>overall</td><td>
<table>
<tr><th>true classes/hypothesized classes</th><th>artificial laugh</th><th>non-artificial laugh</th></tr>
<tr><td>artificial laugh</td><td>98597 (tp)</td><td>18715 (fn)</td></tr>
<tr><td>non-artificial laugh</td><td>9074 (fp)</td><td>293388 (tn)</td></tr>
</table>
precision = 0.91572
recall = 0.84047
accuracy = 0.9338
</td></tr>
</table>

Πίνακας 6.5: Frames classification results

6.3 Διαχωρισμός χιουμοριστικών/μη χιουμοριστικών υποτίτλων

Από την προηγούμενη ενότητα ξέρουμε για όλα τα frames κάθε επεισοδίου, σε ποια κατηγορία ανήκουν. Τώρα από το επίπεδο των frames, θα περάσουμε στο επίπεδο διαστημάτων χρόνου. Για να γίνει αυτό βρίσκουμε τα σημεία αλλαγής κατηγορίας μέσα στα διανύσματα ετικετών που δημιουργήσαμε για τα 23 επεισόδια στην προηγούμενη ενότητα. Πολλαπλασιάζοντας τα σημεία αυτά με την περίοδο του frame, βρίσκουμε τα διαστήματα χρόνου μηχανικού γέλιου/μη μηχανικού γέλιου. Ας δείξουμε με το παρακάτω παράδειγμα τον τρόπο εξαγωγής των διαστημάτων χρόνου:

[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 ...]

Θυμίζουμε ότι με την ετικέτα "1"("0") συμβολίζουμε την κατηγορία μηχανικού γέλιου (μη μηχανικού γέλιου). Βλέπουμε ότι τα σημεία αλλαγών είναι στην θέση 18 και στην θέση 30. Οπότε λαμβάνοντας υπόψιν ότι η περίοδος του frame είναι 0.01 sec, έχουμε τα ακόλουθα διαστήματα χρόνου:
Από 0 sec εως 0.17 sec το οποίο χαρακτηρίζεται ως διάστημα μη μηχανικού γέλιου.
Από 0.18 sec εως 0.29 sec το οποίο χαρακτηρίζεται ως διάστημα μηχανικού γέλιου.

Με βάση την παραπάνω διαδικασία δημιουργούμε 23 πίνακες διαστημάτων χρόνου - κατηγοριών της μορφής:

Number of audio segment	Start time (sec)	End time (sec)	Category (1 για διαστήματα artificial laugh και 0 για τα λοιπά διαστήματα)
1	0	11.287	0
2	11.287	13.323	1
⋮	⋮	⋮	⋮
N	...	...	...

Όπως αναφέραμε στην ενότητα (6.1), έχουμε 23 αρχεία που περιέχουν τους υπότιτλους με τους χρόνους τους, και είναι της μορφής:

1

00:00:02,235 −− > 00:00:04,726

Phoebe found out about this lady...

2

00:00:04,938 −− > 00:00:09,034

...who knew her parents,

and I don't know what happened.

3

00:00:09,309 −− > 00:00:10,901

I'm your mother.

...

Για κάθε ένα από τα παραπάνω αρχεία δημιουργούμε πίνακες διαστημάτων χρόνου και υποτίτλων, της μορφής των πινάκων 6.6, 6.7. Με S συμβολίζουμε το πλήθος των υποτίτλων του επεισοδίου.

Number of subtitle	Start time(sec)	End time(sec)
1	2.235	4.726
2	4.938	9.034
3	9.309	10.901
⋮	⋮	⋮
S	...	...

Πίνακας 6.6: Times of subtitles

Number of subtitle	Subtitles
1	Phoebe found out about this lady...
2	...who knew her parents, and I don't know what happened.
3	I'm your mother.
⋮	⋮
S	...

Πίνακας 6.7: Subtitles

Εξετάζοντας τους πίνακες διαστημάτων χρόνου - κατηγοριών και τους πίνακες με τους χρόνους των υποτίτλων, επιλέγουμε ως χιουμοριστικούς, τους υπότιτλους οι οποίοι βρίσκονται αμέσως πριν από ένα διάστημα χρόνου μηχανικού γέλιου. Οι υπόλοιποι υπότιτλοι θεωρούνται σαν μη χιουμοριστικοί. Ο αλγόριθμος ο οποίος προσδιορίζει τους χιουμοριστικούς/μη χιουμοριστικούς υπότιτλους είναι ο ακόλουθος:

Είσοδοι: Ο πίνακας των διαστημάτων χρόνου - κατηγοριών, και ο πίνακας με τους χρόνους των υποτίτλων.
Έξοδος : Ένα διάνυσμα με μήκος ίσο με των αριθμό τον υποτίτλων, που περιέχει ετικέτες "1" ("0"), που περιγράφουν αν ο υπότιτλος είναι χιουμοριστικός (μη χιουμοριστικός).

1: Αρχικοποίηση του διανύσματος εξόδου με μηδενικά.

2: Για κάθε διάστημα μηχανικού γέλιου, βρες την τελευταία θέση στον πίνακα χρόνων υποτίτλων για την οποία ισχύει:

α) Ο χρόνος έναρξης του υπότιτλου είναι μικρότερος από τον χρόνο έναρξης του μηχανικού γέλιου.

β) Ο χρόνος λήξης του υπότιτλου να μην έχει χρονική διαφορά μεγαλύτερη από 1 sec, με τον χρόνο έναρξης του μηχανικού γέλιου.
Ανάθεσε σε αυτή τη θέση στο διάνυσμα εξόδου την ετικέτα "1".

Σημείωση: Επειδή υπάρχει υψηλή αβεβαιότητα για την κατηγορία των υποτίτλων οι οποίοι βρίσκονται αμέσως πριν από τους χιουμοριστικούς, υπό την έννοια ότι υπάρχει πιθανότητα να συμμετέχουν και αυτοί στο χιούμορ, τους απορρίπτουμε.

Εκτελώντας τον παραπάνω αλγόριθμο για τα 23 επεισόδια, δημιουργούμε 23 διανύσματα, το κάθε ένα από τα οποία περιέχει ετικέτες (χιουμοριστικός/μη χιουμοριστικός) για όλους τους υπότιτλους του αντίστοιχου επεισοδίου. Στον επόμενο πίνακα παραθέτουμε τα στατιστικά στοιχεία (για τα 23 επεισόδια) που προέκυψαν από την παραπάνω διαδικασία.

Number of subtitles	8036
Humorous	2820 (35.1%)
Non-humorous	3325 (41.4%)
Rejected	1891 (23.5%)

Πίνακας 6.8: Subtitles Statistics

Για να εκτιμήσουμε την απόδοση του ταξινομητή, πρέπει να τρέξουμε τον αλγόριθμο ταξινόμησης των υποτίτλων με είσοδο τους πίνακες διαστημάτων χρόνου - κατηγοριών που προέκυψαν από το χειροκίνητο labeling, για τα τρία πρώτα επεισόδια (ενότητα 6.1). Τα διανύσματα που θα επιστρέψει ο αλγόριθμος αντιπροσωπεύουν τις σωστές αποφάσεις για την κατηγορία κάθε υπότιτλου από τα τρία πρώτα επεισόδια. Συγκρίνοντας αυτά τα διανύσματα με τα αντίστοιχα διανύσματα που προέκυψαν από την διαδικασία του testing, μπορούμε να δούμε ποιοι υπότιτλοι ταξινομήθηκαν

σωστά. Έστω tp (fn) ο αριθμός των χιουμοριστικών υποτίτλων που ταξινομήθηκαν σωστά (λάθος). Αντίστοιχα συμβολίζουμε με tn (fp) τον αριθμό των μη χιουμοριστικών υποτίτλων που ταξινομήθηκαν σωστά (λάθος). Στον πίνακα 6.9 παραθέτουμε τα αποτελέσματα.

<table>
<tr><th>Episodes</th><th>misclassification tables</th></tr>
<tr><td>3</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>119 (tp)</td><td>15 (fn)</td></tr>
<tr><td>non humorous</td><td>4 (fp)</td><td>180 (tn)</td></tr>
</table>
precision = 0.96748
recall = 0.88806
accuracy = 0.94025
</td></tr>
<tr><td>2</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>152 (tp)</td><td>11 (fn)</td></tr>
<tr><td>non humorous</td><td>1 (fp)</td><td>168 (tn)</td></tr>
</table>
precision = 0.99346
recall = 0.93252
accuracy = 0.96386
</td></tr>
<tr><td>1</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>119 (tp)</td><td>4 (fn)</td></tr>
<tr><td>non humorous</td><td>8 (fp)</td><td>189 (tn)</td></tr>
</table>
precision = 0.93701
recall = 0.96748
accuracy = 0.9625
</td></tr>
<tr><td>overall</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>390 (tp)</td><td>30 (fn)</td></tr>
<tr><td>non humorous</td><td>13 (fp)</td><td>537 (tn)</td></tr>
</table>
precision = 0.96774
recall = 0.92857
accuracy = 0.95567
</td></tr>
</table>

Πίνακας 6.9: Subtitles classification results

Κεφάλαιο 7

Αυτόματη αναγνώριση χιουμοριστικών και μη χιουμοριστικών υποτίτλων

Από την προηγούμενη φάση της εργασίας μας (κεφάλαιο 6), έχουμε ξεχωρίσει τους χιουμοριστικούς από τους μη χιουμοριστικούς υπότιτλους για τα 23 επεισόδια της 4ης σεζόν της σειράς Friends. Στόχος σε αυτή τη φάση της εργασίας μας είναι να γίνει αυτόματη αναγνώριση των υποτίτλων (αν είναι χιουμοριστικοί ή όχι), με βάση γλωσσικά χαρακτηριστικά. Για τον σκοπό αυτό θα χρησιμοποιήσουμε τρεις μεθόδους:

- N-gram πιθανοτικό μοντέλο
- Μοντέλο διανυσματικού χώρου
- Λανθάνουσα σημασιολογική ανάλυση.

7.1 Τα δεδομένα μας - Χωρισμός σε training και testing sets

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, έχουμε δημιουργήσει μια συλλογή από χιουμοριστικούς και μη χιουμοριστικούς υπότιτλους. Στον παρακάτω πίνακα

φαίνεται ο συνολικός αριθμός των υποτίτλων, καθώς και η κατανομή τους σε χιουμοριστικούς και μη χιουμοριστικούς:

Αριθμός υποτίτλων	6145
χιουμοριστικοί	2820 (45.9%)
μη χιουμοριστικοί	3325 (54.1%)

Πίνακας 7.1: Lexical Corpus

Μία συνήθης τακτική είναι να αφαιρούνται τα **stop words** από την συλλογή των κειμένων. Stop words είναι λέξεις οι οποίες δεν προσφέρουν χρήσιμη γλωσσολογική πληροφορία (άρθρα, προθέσεις,..). Τα stop words υπάρχουν σε μεγάλο αριθμό μέσα στα κείμενα, και εφόσον δεν προσφέρουν κάποια πληροφορία διακριτοποίησης μεταξύ των κειμένων, προσθέτουν μόνο θόρυβο. Έτσι λοιπόν, αφαιρώντας τα stop words μειώνουμε τον θόρυβο και κάνουμε τα δεδομένα μας καταλληλότερα για την μετέπειτα διαδικασία. Επίσης αφαιρούνται μη λεκτικά σύμβολα, όπως σημεία στίξης και αριθμοί. Τέλος μετατρέπουμε όλους τους χαρακτήρες σε κεφαλαίους, έτσι ώστε όλοι οι χαρακτήρες να είναι σε μία τυποποιημένη μορφή (case folding). Μετά από την διαδικασία αυτή, χωρίζουμε τα δεδομένα μας σε training και testing sets. Για την δημιουργία των training και testing δεδομένων, χρησιμοποιούμε δύο τακτικές :

1. Επιλέγουμε το 30% των δεδομένων μας (το οποίο αντιστοιχεί στους υπότιτλους από 7 επεισόδια) ως testing set και το υπόλοιπο 70% (το οποίο αντιστοιχεί στους υπότιτλους από τα υπόλοιπα 16 επεισόδια) ως training set. Στον πίνακα 7.2 δείχνουμε την κατανομή σε χιουμοριστικούς και μη χιουμοριστικούς υπότιτλους στο training και testing set.

2. Με τον αλγόριθμο του cross-validation (leave one out) κατασκευάζουμε 23 training sets με τα αντίστοιχα 23 testing sets. Συγκεκριμένα κάθε testing set αντιστοιχεί σε ένα επεισόδιο (από το 1 έως το 23) και το αντίστοιχο training set αντιστοιχεί στα υπόλοιπα 22 επεισόδια. Στον πίνακα 7.3 φαίνονται αναλυτικά τα training με αντίστοιχα testing sets.

	# humorous subtitles	# non humorous subtitles
training set	1988	2400
testing set	832	925

Πίνακας 7.2: Training/testing sets in "70%-30%" experiment

training/testing sets	# humorous subtitles	# non humorous subtitles
training set 1	2712	3181
testing set 1	108	144
training set 2	2683	3210
testing set 2	137	115
training set 3	2704	3195
testing set 3	116	130
training set 4	2704	3192
testing set 4	116	133
training set 5	2690	3201
testing set 5	130	124
training set 6	2705	3203
testing set 6	115	122
training set 7	2710	3168
testing set 7	110	157
training set 8	2712	3149
testing set 8	108	176
training set 9	2697	3197
testing set 9	123	128
training set 10	2700	3193
testing set 10	120	132
training set 11	2709	3150
testing set 11	111	175
training set 12	2691	3208
testing set 12	129	117
training set 13	2700	3210
testing set 13	120	115

training/testing sets	# humorous subtitles	# non humorous subtitles
training set 14	2693	3174
testing set 14	127	151
training set 15	2706	3185
testing set 15	114	140
training set 16	2681	3215
testing set 16	139	110
training set 17	2706	3183
testing set 17	114	142
training set 18	2696	3202
testing set 18	124	123
training set 19	2708	3178
testing set 19	112	147
training set 20	2698	3201
testing set 20	122	124
training set 21	2741	3145
testing set 21	79	180
training set 22	2717	3174
testing set 22	103	151
training set 23	2577	3036
testing set 23	243	289

Πίνακας 7.3: Training/testing sets in cross validation experiment

7.2 N-gram πιθανοτικό μοντέλο

Στην ενότητα αυτή θα προσπαθήσουμε με την βοήθεια του εργαλείου CMU, να κατασκευάσουμε ένα ταξινομητή που θα αποφασίζει αν ένας υπότιτλος είναι χιουμοριστικός ή όχι. Συγκεκριμένα με βάση το εκάστοτε training set (ενότητα 7.1) δημιουργούμε γλωσσικά μοντέλα για τις κατηγορίες των χιουμοριστικών υποτίτλων (w_1), και των μη χιουμοριστικών υποτίτλων (w_2). Τα μοντέλα που θα χρησιμοποιήσουμε είναι unigram, bigram, trigram. Υπολογίζονται οι a-priori πιθανότητες κάθε κατηγορίας με βάση το training set:

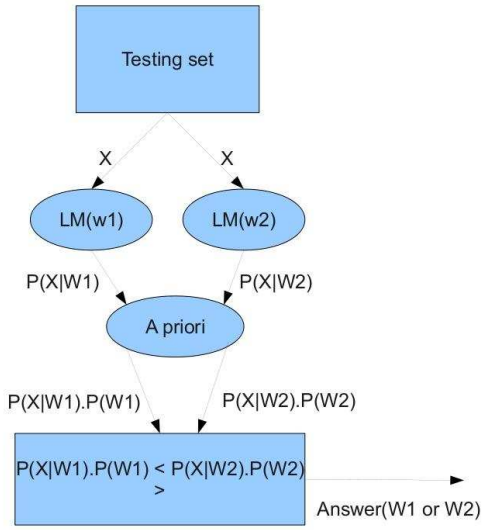
$$P(\omega_1) = \frac{\textbf{\# Χιουμοριστικών υποτίτλων στο training set}}{\textbf{\# Χιουμοριστικών και μη χιουμοριστικών υποτίτλων στο training set}} \quad (7.1)$$

$$P(\omega_2) = \frac{\textbf{\# Μη χιουμοριστικών υποτίτλων στο training set}}{\textbf{\# Χιουμοριστικών και μη χιουμοριστικών υποτίτλων στο training set}} \quad (7.2)$$

Για κάθε υπότιτλο x από το testing set, υπολογίζεται η πιθανότητα του, σύμφωνα με το μοντέλο για τους χιουμοριστικούς υπότιτλους ($P(x|\omega_1)$), και σύμφωνα με το μοντέλο για τους μη χιουμοριστικούς υπότιτλους ($P(x|\omega_2)$). Ο υπότιτλος ταξινομείται στην κατηγορία η οποία δίνει την μεγαλύτερη πιθανότητα:

$$\widehat{\omega} = \arg\max_{\omega_i} P(\omega_i|x) = \arg\max_{\omega_i}\{P(x|\omega_i)\ .\ P(\omega_i)\}\ ,\ i = 1, 2 \qquad (7.3)$$

Τέλος ξέροντας από την ενότητα 6.3 σε ποια κατηγορία ανήκει κάθε υπότιτλος, μπορούμε να δούμε αν ταξινομήθηκε σωστά. Στο σχήμα 7.1 δείχνουμε τον ταξινομητή μας.



Σχήμα 7.1: Εικόνα του ταξινομητή

7.2.1 Πειραματική διαδικασία

Με βάση το εκάστοτε training set (ενότητα 7.1) και την πληροφορία της κατηγορίας στην οποία ανήκει κάθε υπότιτλος (από την ενότητα 6.3), διαχωρίζουμε τους υπότιτλους σε δύο αρχεία. Το *train_humorous.txt* το οποίο περιέχει όλους τους

χιουμοριστικούς υπότιτλους από το training set και το *train_NONhumorous.txt* το οποίο περιέχει όλους τους μη χιουμοριστικούς υπότιτλους από το training set.

Για την κατασκευή των γλωσσικών μοντέλων, ακολουθούμε τα εξής βήματα:

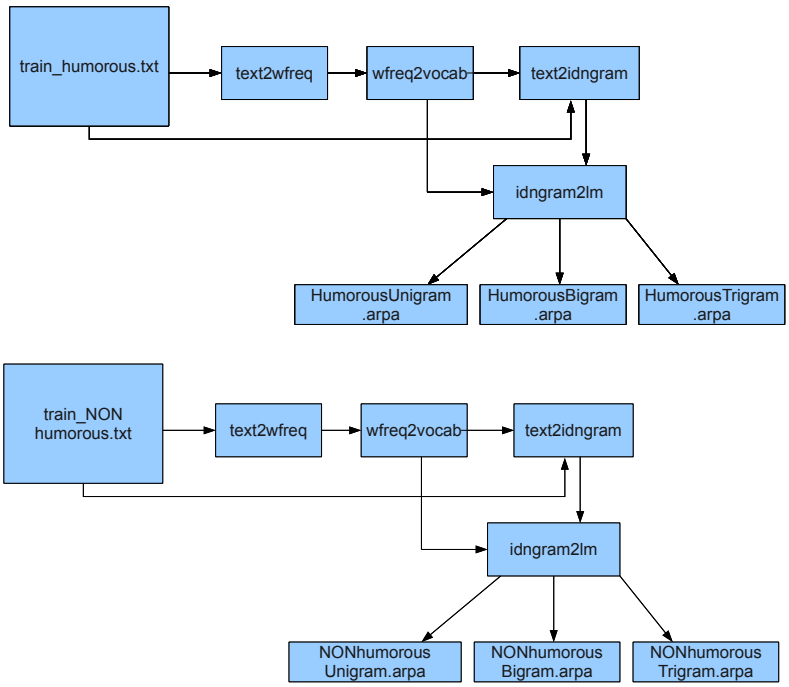
1. Με την συνάρτηση text2wfreq και το αρχείο *train_humorous.txt* δημιουργούμε ένα αρχείο humorous.wfreq. Το αρχείο αυτό είναι σε -ascii μορφή και περιέχει μια λίστα με τις λέξεις και τον αριθμό των εμφανίσεων τους, στο κείμενο *train_humorous.txt*. Δημιουργούμε το αντίστοιχο αρχείο για το *train_NONhumorous.txt*, το οποίο είναι το NONhumorous.wfreq.

2. Χρησιμοποιώντας την συνάρτηση wfreq2vocab και το αρχείο humorous.wfreq, δημιουργούμε ένα αρχείο humorous.vocab. Το humorous.vocab είναι ένα -ascii αρχείο το οποίο περιέχει το λεξιλόγιο της κατηγορίας humorous. Κατασκευάζουμε το αντίστοιχο αρχείο για την κατηγορία non humorous, το οποίο είναι το NONhumorous.vocab.

3. Με τα αρχεία humorous.vocab, *train_humorous.txt*, και την συνάρτηση text2 idngram δημιουργούμε τρία αρχεία: Τα humorous.id1gram, humorous.id2gram, humorous.id3gram, τα οποία περιέχουν αντιστοίχως, τα unigrams, unigrams/ bigrams, unigrams/bigrams/trigrams, του κειμένου *train_humorous.txt* καθώς και τον αριθμό των εμφανίσεων τους μέσα στο κείμενο. Τα αντίστοιχα για την κατηγορία non humorous είναι τα NONhumorous.id1gram, NONhumorous .id2gram, NONhumorous.id3gram.

4. Τέλος με την συνάρτηση idngram2lm και τα αρχεία humorous.id1gram, humorous.id2gram, humorous.id3gram, humorous.vocab κατασκευάζουμε τρία γλωσσικά μοντέλα για την κατηγορία humorous. Τα humorousUnigram.arpa, humorousBigram.arpa, humorousTrigram.arpa, τα οποία περιέχουν αντιστοίχως όλα τα unigrams, unigrams/bigrams, unigrams/bigrams/trigrams και τις αντίστοιχες πιθανότητες τους στο κείμενο *train_humorous.txt*. Δημιουργούμε τα αντίστοιχα αρχεία για την κατηγορία non humorous (NONhumorous Unigram.arpa, NONhumorousBigram.arpa, NONhumorousTrigram.arpa).

Σημείωση: Για την κατασκευή των γλωσσικών μοντέλων, χρησιμοποιήθηκε Good Turing discounting. Επίσης τα μοντέλα είναι "open vocabulary", επιτρέποντας σε λέξεις εκτός λεξιλογίου να εμφανιστούν. Όλες οι λέξεις οι οποίες δεν υπάρχουν στο λεξιλόγιο αντιστοιχίζονται στο ίδιο σύμβολο. Η παραπάνω διαδικασία παρουσιάζεται στο σχήμα 7.2.



Σχήμα 7.2: Δημιουργία γλωσσικών μοντέλων

Έχοντας δημιουργήσει τα γλωσσικά μοντέλα μας (για την χιουμοριστική κατηγορία (c_1) και την μη χιουμοριστική κατηγορία (c_2)), είμαστε έτοιμοι με την βοήθεια της συνάρτησης evallm, να υπολογίσουμε την πιθανότητα κάθε υπότιτλου x από το testing set, σύμφωνα με τα γλωσσικά μοντέλα για την χιουμοριστική κατηγορία

($P(x|c_1)$) και μη χιουμοριστική κατηγορία ($P(x|c_2)$). Συγκεκριμένα η evallm δέχεται σαν ορίσματα το γλωσσικό μοντέλο και το κείμενο του υπότιτλου και επιστρέφει για κάθε λέξη στο κείμενο του υπότιτλου, την πιθανότητα να ανήκει στην κατηγορία που περιγράφεται από το γλωσσικό μοντέλο. Πολλαπλασιάζοντας τις πιθανότητες των λέξεων του υπότιτλου παράγεται ένα γινόμενο πιθανοτήτων. Το γινόμενο αυτό πολλαπλασιάζεται και με την a-priori πιθανότητα της κατηγορίας. Λογαριθμίζουμε το γινόμενο των πιθανοτήτων, αφού μπορεί να είναι πολύ μικρός αριθμός και να έχουμε προβλήματα numerical underflow. Ο υπότιτλος ταξινομείται στην κατηγορία η οποία δίνει την μεγαλύτερη πιθανότητα. Αν θεωρήσουμε ότι ο αριθμός των λέξεων στο κείμενο του υπότιτλου είναι n $(w_1, w_2, w_3, ..., w_n)$, μπορούμε να γράψουμε τα παραπάνω τα παραπάνω με μαθηματικό φορμαλισμό ως εξής:

1. Περίπτωση unigram γλωσσικού μοντέλου:

$$\begin{aligned} \widehat{c} &= \arg\max_{c_i} P(c_i|x) = \arg\max_{c_i}\{P(x|c_i)\,.\,P(c_i)\} = \\ &= \arg\max_{c_i}\{P(c_i)\,.\,\prod_{j=1}^{n} P(w_j|c_i)\} = \\ &= \arg\max_{c_i}\{\log(P(c_i)) + \sum_{j=1}^{n}\log(P(w_j|c_i))\},\ i=1,2 \end{aligned} \tag{7.4}$$

2. Όμοια στην περίπτωση του bigram γλωσσικού μοντέλου:

$$\widehat{c} = \arg\max_{c_i}\{\log(P(c_i)) + \sum_{j=1}^{n}\log(P(w_j|w_{j-1},\ c_i))\},\ i=1,2 \tag{7.5}$$

3. Τέλος στην περίπτωση trigram γλωσσικού μοντέλου:

$$\widehat{c} = \arg\max_{c_i}\{\log(P(c_i)) + \sum_{j=1}^{n}\log(P(w_j|w_{j-1},\ w_{j-2},\ c_i))\},\ i=1,2 \tag{7.6}$$

Με το παρακάτω παράδειγμα (στο οποίο τα μοντέλα έχουν εκπαιδευτεί με το 70% των δεδομένων (ενότητα 7.1 - περίπτωση 1)) θα δείξουμε πως υπολογίζεται η

πιθανότητα που έχει ο υπότιτλος με κάθε κατηγορία:
Ο υπότιτλος x προς ταξινόμηση είναι ο εξής:
PHOEBE FOUND OUT ABOUT THIS LADY
Ακολουθούν τα παρακάτω βήματα:

- Υπολογισμός των a-priori πιθανοτήτων κάθε κατηγορίας (σύμφωνα με τους τύπους 7.1, 7.2):
  Για την κατηγορία των χιουμοριστικών υποτίτλων (c_1):
  $P(c_1) = 0.453 \Rightarrow log(P(c_1)) = -0.344$
  Για την κατηγορία των μη χιουμοριστικών υποτίτλων (c_2):
  $P(c_2) = 0.547 \Rightarrow log(P(c_2)) = -0.262$

- Με την βοήθεια της evallm, υπολογίζονται οι πιθανότητες που έχει ο υπότιτλος x, αναφορικά με κάθε κατηγορία ($P(x|c_1),\ P(x|c_2)$).

Στο παράδειγμα αυτό θα αναφερθούμε στην περίπτωση των trigram μοντέλων.
Έτσι είσοδο το humorousTrigram.arpa, η evallm παράγει το εξής κείμενο:

$P(PHOEBE \mid)$ = 0.00066359 logprob = -3.178100 *bo_case* = 1

$P(FOUND \mid PHOEBE)$ = 0.000105245 logprob = -3.977800 *bo_case* = 2-1

$P(OUT \mid PHOEBE\ FOUND)$ = 0.00206443 logprob = -2.685200 *bo_case* = 3x2-1

$P(ABOUT \mid FOUND\ OUT)$ = 0.000888383 logprob = -3.051400 *bo_case* = 3x2-1

$P(THIS \mid OUT\ ABOUT)$ = 0.00640767 logprob = -2.193300 *bo_case* = 3x2

$P(LADY \mid ABOUT\ THIS)$ = 9.62942e-05 logprob = -4.016400 *bo_case* = 3-2-1

Με είσοδο το NONhumorousTrigram.arpa, η evallm παράγει το εξής κείμενο:

$P(PHOEBE \mid)$ = 0.000492493 logprob = -3.307600 *bo_case* = 1

$P(FOUND \mid PHOEBE)$ = 0.000131704 logprob = -3.880400 *bo_case* = 2-1

$P(OUT \mid PHOEBE\ FOUND)$ = 0.16233 logprob = -0.789600 *bo_case* = 3x2

$P(ABOUT \mid FOUND\ OUT)$ = 0.0014873 logprob = -2.827600 *bo_case* = 3-2-1

$P(THIS \mid OUT\ ABOUT)$ = 0.0514991 logprob = -1.288200 *bo_case* = 3x2

$P(LADY \mid ABOUT\ THIS)$ = 3.23743e-05 logprob = -4.489800 *bo_case* = 3-2-1

Εφόσον το μοντέλα μας είναι trigram, ο υπότιτλος ταξινομείται σύμφωνα με τον τύπο 7.6:
$Score_{c_1} = log(P(c_1)) + log(P(x|c_1)) = -0.344 - 3.1781 - 3.9778 - 2.6852 - 3.0514 - 2.1933 - 4.0164 = -19.4462$

$Score_{c_2} = log(P(c_2)) + log(P(x|c_2)) = -0.262 - 3.3076 - 3.8804 - 0.7896 - 2.8276 - 1.2882 - 4.4898 = -16.8452$

Αφού $Score_{c_2} > Score_{c_1}$, ο υπότιτλος ταξινομείται ως μη χιουμοριστικός.

Αναφορές

Ο αναγνώστης μπορεί να ανατρέξει στην τεκμηρίωση του εργαλείου CMU [13], για επιπλέον πληροφορίες σχετικά με την λειτουργικότητα του.

7.3 Το μοντέλο διανυσματικού χώρου

7.3.1 Εξαγωγή των όρων από τα κείμενα και κατασκευή του πίνακα όρων-κειμένων

Με βάση το εκάστοτε training set (ενότητα 7.1), και την πληροφορία της κατηγορίας στην οποία ανήκει κάθε υπότιτλος (από την ενότητα 6.3), διαχωρίζουμε τους υπότιτλους σε δύο αρχεία. Το *train_humorous.txt* το οποίο περιέχει όλους τους χιουμοριστικούς υπότιτλους από το training set και το *train_NONhumorous.txt* το οποίο περιέχει όλους τους μη χιουμοριστικούς υπότιτλους από το training set. Αυτά τα δύο κείμενα αποτελούν τα έγγραφα της συλλογής μας. Έπειτα από την δημιουργία των παραπάνω κειμένων, γίνεται εξαγωγή των όρων από κάθε κείμενο. Επίσης γίνεται εξαγωγή των όρων από κάθε υπότιτλο του αντίστοιχου testing set (ενότητα 7.1), ώστε να δημιουργηθούν τα διανύσματα των ερωτήσεων πάνω στην συλλογή. Το πλήθος των διανυσμάτων αυτών είναι είναι ίσο με τον αριθμό των υποτίτλων του testing set. Αναλόγως με το μοντέλο μας, ακολουθούμε τις επόμενες επιλογές σχετικά με την εξαγωγή των όρων:

1. Εξαγωγή unigram terms
2. Εξαγωγή unigram/bigram terms
3. Εξαγωγή unigram/bigram/trigram terms

Στον πίνακα 7.4 φαίνονται οι όροι που έχουν εξαχθεί με βάση το training set του πίνακα 7.2 (ενότητα 7.1-περίπτωση 1). Στον πίνακα 7.5 παρουσιάζουμε τους όρους που έχουν εξαχθεί από τα training sets του cross validation πειράματος.

unigrams	bigrams	trigrams
2895	11293	13577

Πίνακας 7.4: Training features in "70%-30%" experiment

training set number	unigrams	bigrams	trigrams
1	3473	14274	17886
2	3448	14193	17793
3	3435	14149	17757
4	3430	14141	17791
5	3460	14175	17797
6	3449	14207	17793
7	3456	14160	17743
8	3467	14128	17603
9	3414	14100	17658
10	3450	14160	17755
11	3451	14130	17657
12	3406	14118	17797
13	3451	14180	17772
14	3439	14090	17696
15	3464	14182	17821
16	3451	14120	17839
17	3455	14139	17674
18	3440	14155	17774
19	3478	14252	17917
20	3495	14296	17908
21	3472	14265	17911
22	3461	14157	17724
23	3365	13710	17036

Πίνακας 7.5: Training features in cross validation experiment

Κάθε έγγραφο της συλλογής μας (όπως και κάθε ερώτηση πάνω στην συλλογή) αναπαριστάται ως ένα διάνυσμα από όρους δεικτοδότησης (index terms). Όπως αναφέρθηκε ανάλογα με το μοντέλο μας οι όροι μπορεί να είναι unigrams, unigrams/bigrams, unigrams/bigrams/trigrams. Το σύνολο όλων των μοναδικών όρων της συλλογής μας αποτελεί το λεξιλόγιο (vocabulary) της συλλογής μας. Αυτοί οι όροι σχηματίζουν ένα διανυσματικό χώρο, διάστασης ίσης με το μέγεθος του λεξιλογίου. Αν θεωρήσουμε ότι το πλήθος των διαφορετικών όρων της συλλογής μας είναι N, τότε τα κείμενα και οι ερωτήσεις εκφράζονται ως N-διάστατα διανύσματα:

$\underline{d}_j(w_{1j}, w_{2j}, ..., w_{Nj}),\ j = 1,\ 2$

$\underline{q}(w^q_{11}, w^q_{21}, ..., w^q_{N1})$

Το j=1 αφορά το κείμενο με τους χιουμοριστικούς υπότιτλους, ενώ το j=2 αφορά το κείμενο με τους μη χιουμοριστικούς. Όπως φαίνεται από τον παραπάνω συμβολισμό, σε κάθε όρο i μέσα σε ένα κείμενο j (ερώτηση q), ανατίθεται ένα βάρος w_{ij} (w_{i1}^q). Το βάρος αυτό υποδηλώνει πόσο σημαντικός είναι ο όρος i στο κείμενο j (ερώτηση q), στην διακριτοποίηση του από τα άλλα κείμενα της συλλογής. Η ανάθεση των βαρών στον πίνακα όρων-κειμένων (term document matrix) και στα διανύσματα των ερωτήσεων αναλύεται στην επόμενη ενότητα. Συνοψίζοντας λοιπόν μπορούμε να πούμε ότι η συλλογή μας έχει αναπαρασταθεί από ένα πίνακα όρων-κειμένων, οι γραμμές του οποίου αναπαριστούν τους όρους και οι στήλες τα κείμενα. Σε κάθε θέση (i,j) του πίνακα, υπάρχει το βάρος του όρου i στο κείμενο j. Όπως προαναφέραμε αυτό το βάρος φανερώνει την σημαντικότητα του όρου i στο κείμενο j. Αν δεν υπάρχει ο όρος αυτός μέσα στο κείμενο τότε του ανατίθεται μηδενικό βάρος. Ο πίνακας όρων-κειμένων του μοντέλου μας φαίνεται στον πίνακα 7.6. Το document 1 αφορά τους χιουμοριστικούς υπότιτλους από το training set (*train_humorous.txt*), και το document 2 τους μη χιουμοριστικούς από το training set (*train_NONhumorous.txt*). Η μορφή του διανύσματος μίας ερώτησης πάνω στην συλλογή φαίνεται στον πίνακα 7.7.

	document 1	document 2
term 1	w_{11}	w_{12}
term 2	w_{21}	w_{22}
⋮	⋮	⋮
term N	w_{N1}	w_{N2}

Πίνακας 7.6: Term document matrix

	query
term 1	w_{11}^q
term 2	w_{21}^q
⋮	⋮
term N	w_{N1}^q

Πίνακας 7.7: Query vector

7.3.2 Ανάθεση βαρών στον πίνακα όρων-κειμένων και στα διανύσματα των ερωτήσεων

Για για να αναθέσουμε τα βάρη w_{ij} στον πίνακα όρων-κειμένων του μοντέλου μας χρησιμοποιήσαμε το σχήμα ανάθεσης βαρών tf-idf, όπως ορίστηκε στην ενότητα 3.1. Θυμίζουμε ότι: $w_{ij} = tf_{ij}\ idf_i$, όπου $tf_{ij} = \frac{f_{ij}}{\max_l f_{lj}}$ και $idf_i = \log \frac{1+D}{n_i}$. Το f_{ij} είναι ο αριθμός των εμφανίσεων του όρου i στο κείμενο j. Το D είναι το πλήθος των κειμένων της συλλογής, ενώ το n_i είναι το πλήθος των κειμένων στα οποία εμφανίζεται ο όρος i. Στο δικό μας μοντέλο έχουμε δύο κείμενα (το πρώτο αφορά τους χιουμοριστικούς υπότιτλους και το δεύτερο τους μη χιουμοριστικούς). Έτσι εάν ένας όρος i εμφανίζεται και στα δύο, τότε θα παίρνει τιμή $idf_i = \log \frac{3}{2}$. Διαφορετικά (αν εμφανίζεται μόνο σε ένα κείμενο) θα παίρνει τιμή $idf_i = \log 3$. Πριμοδοτούμε δηλαδή τους όρους οι οποίοι εμφανίζονται μόνο σε ένα κείμενο, αφού χαρακτηρίζουν καλύτερα την κλάση των εγγράφων μας. Μια παρατήρηση στον τύπο που δίνει το idf, είναι ότι προσθέτουμε το 1 στον αριθμό των εγγράφων (D). Αυτό γίνεται διότι διαφορετικά όλοι οι όροι που υπάρχουν και στα δύο έγγραφα θα είχαν βάρος μηδέν.

Για την ανάθεση των βαρών στα διανύσματα των ερωτήσεων, ακολουθήσαμε διαφορετική τακτική. Έστω ένας υπότιτλος j προς ταξινόμηση. Τότε όπως αναφέραμε και στην προηγούμενη ενότητα, γίνεται εξαγωγή των όρων του υπότιτλου και δημιουργείται ένα διάνυσμα ερώτησης j. Σε κάθε όρο του υπότιτλου που υπάρχει στην συλλογή μας (άρα και στον πίνακα όρων-κειμένων) ανατίθεται (στην κατάλληλη θέση στο διάνυσμα της ερώτησης) βάρος το οποίο ισούται με τον αριθμό των εμφανίσεων του όρου στο κείμενο του υπότιτλου. Στους υπόλοιπους όρους του διανύσματος της ερώτησης ανατίθεται μηδενικό βάρος.

7.3.3 Μέτρηση ομοιότητας κειμένου-ερώτησης

Έχοντας δημιουργήσει τον πίνακα όρων-κειμένων του μοντέλου μας, δεχόμαστε ερωτήσεις. Μια ερώτηση σχετίζεται με ένα υπότιτλο, ο οποίος περιμένει να ταξινομηθεί στην κατηγορία των χιουμοριστικών ή των μη χιουμοριστικών υποτίτλων. Το κείμενο του υπότιτλου αντιπροσωπεύεται από το διάνυσμα της ερώτησης. Για

να βρούμε ποιο διάνυσμα κειμένου, είναι περισσότερο όμοιο με το διάνυσμα της ερώτησης, χρησιμοποιούμε την μετρική cosine similarity. Έστω $\mathbf{d}$ το διάνυσμα ενός κειμένου και $\mathbf{q}$ το διάνυσμα της ερώτησης. Τότε η ομοιότητα μεταξύ του διανύσματος του κειμένου $\mathbf{d}$ και της ερώτησης $\mathbf{q}$ ορίζεται ως:

$$sim(\mathbf{d}, \mathbf{q}) = \frac{\mathbf{d}.\mathbf{q}}{|\mathbf{d}||\mathbf{q}|} = \frac{\sum_{1 \leq i \leq N} d_i q_i}{\sqrt{\sum_{1 \leq i \leq N} d_i^2} \sqrt{\sum_{1 \leq i \leq N} q_i^2}}$$

Για κάθε διάνυσμα ερώτησης, βρίσκεται η ομοιότητα με το διάνυσμα κειμένου των χιουμοριστικών και το διάνυσμα κειμένου των μη χιουμοριστικών υποτίτλων. Ο υπότιτλος ταξινομείται στην κατηγορία της οποίας το διάνυσμα κειμένου είναι περισσότερο όμοιο με το διάνυσμα της ερώτησης. Ξέροντας από την ενότητα 6.3 σε ποια κατηγορία ανήκει κάθε υπότιτλος των testing δεδομένων, μπορούμε να δούμε ποιοι υπότιτλοι ταξινομήθηκαν σωστά.

7.4 Λανθάνουσα σημασιολογική ανάλυση

Το πρώτο βήμα για την υλοποίηση του αλγορίθμου της λανθάνουσας σημασιολογικής ανάλυσης (Latent Semantic Analysis - LSA), ήταν η κατασκευή του πίνακα όρων-κειμένων και των διανυσμάτων των ερωτήσεων, με βάση τα εκάστοτε training/testing sets (ενότητα 7.1). Για να κατασκευάσουμε τον πίνακα και τα διανύσματα, ακολουθήσαμε την διαδικασία που περιγράψαμε στην ενότητα 7.3, οπότε δεν θα την επαναλάβουμε. Το δεύτερο βήμα ήταν να αναπαραστήσουμε τον πίνακα όρων-κειμένων, με την τεχνική της διάσπασης ιδιοτιμών (Singular Value Decomposition - SVD). Έστω A ο πίνακας όρων-κειμένων του μοντέλου μας, διάστασης N x 2. Θυμίζουμε ότι N είναι το πλήθος των όρων μας και ότι το πλήθος των κειμένων μας είναι δύο (το πρώτο αφορά τους χιουμοριστικούς υπότιτλους και το δεύτερο τους μη χιουμοριστικούς). Εφαρμόζοντας SVD στον πίνακα A, διασπάται σε γινόμενο τριών πινάκων ως εξής:

$$A = USV^T$$

Ο U είναι ένας ορθογώνιος πίνακας μεγέθους N x 2, ο οποίος περιέχει τα ιδιοδιανύσματα του AA^T. Ο S είναι διαγώνιος πίνακας διάστασης 2 x 2, όπου 2 είναι η τάξη του πίνακα A. Τα διαγώνια στοιχεία του πίνακα S είναι οι ιδιοτιμές του πίνακα A. Τέλος ο πίνακας V είναι ένας ορθογώνιος πίνακας μεγέθους 2 x 2 , ο οποίος περιέχει τα ιδιοδιανύσματα του A^TA. Παρακάτω δείχνουμε τον πίνακα όρων-κειμένων, το διάνυσμα μίας ερώτησης q, καθώς και την εφαρμογή SVD στον πίνακα A:

$$A : \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ \vdots & \vdots \\ w_{N1} & w_{N2} \end{bmatrix} \quad q : \begin{bmatrix} w_{11}^q \\ w_{21}^q \\ \vdots \\ w_{N1}^q \end{bmatrix}$$

$$A : \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ \vdots & \vdots \\ w_{N1} & w_{N2} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ \vdots & \vdots \\ u_{N1} & u_{N2} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}^T$$

Έπειτα από τα τα βήματα που αναλύσαμε προηγουμένως, είμαστε έτοιμοι να αναπαραστήσουμε το διάνυσμα μιας ερώτησης q, στον διανυσματικό μας χώρο, κάνοντας τον ακόλουθο μετασχηματισμό: $q = q^T U S^{-1}$. Με αυτό το μετασχηματισμό δημιουργείται ένα διάνυσμα 1 x 2, το οποίο αναπαριστά την ερώτηση μας, στον διανυσματικό χώρο του LSA. Οι συντεταγμένες του πρώτου κειμένου (χιουμοριστικών υποτίτλων), βρίσκονται στην πρώτη σειρά του πίνακα V. Αντιστοίχως οι συντεταγμένες του δεύτερου κειμένου (μη χιουμοριστικών υποτίτλων), βρίσκονται στην δεύτερη σειρά του πίνακα V. Οπότε μπορούμε να γράψουμε:

Συντεταγμένες κειμένου χιουμοριστικών υποτίτλων: $\mathbf{d}1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix}$

Συντεταγμένες κειμένου μη χιουμοριστικών υποτίτλων: $\mathbf{d}2 = \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix}$

Συντεταγμένες της ερώτησης στον διανυσματικό χώρο του LSA: $\mathbf{q} = \begin{bmatrix} q_{11} & q_{12} \end{bmatrix}$

Για να δούμε ποιο διάνυσμα κειμένου, είναι περισσότερο όμοιο με το διάνυσμα της ερώτησης, χρησιμοποιούμε την μετρική cosine similarity:

$$sim(\mathbf{d}, \mathbf{q}) = \frac{\sum_{1 \leq i \leq 2} d_{i1}\, q_{1i}}{\sqrt{\sum_{1 \leq i \leq 2} d_{i1}^2} \sqrt{\sum_{1 \leq i \leq 2} q_{1i}^2}}$$

Κλείνοντας να σημειώσουμε ότι ξέρουμε την κατηγορία κάθε υπότιτλου των testing δεδομένων (ενότητα 6.3). Αν ο υπότιτλος που αναπαριστάται από το διάνυσμα της ερώτησης q είναι χιουμοριστικός και $sim(\mathbf{d}1, \mathbf{q}) > sim(\mathbf{d}2, \mathbf{q})$, τότε ταξινομείται σωστά στην κατηγορία των χιουμοριστικών υποτίτλων. Διαφορετικά ταξινομείται λανθασμένα στην κατηγορία των μη χιουμοριστικών υποτίτλων. Όμοια αν ο υπότιτλος είναι μη χιουμοριστικός και $sim(\mathbf{d}2, \mathbf{q}) > sim(\mathbf{d}1, \mathbf{q})$, τότε ταξινομείται σωστά στην κατηγορία των μη χιουμοριστικών υποτίτλων. Διαφορετικά ταξινομείται λανθασμένα στην κατηγορία των χιουμοριστικών υποτίτλων.

7.5 Μελέτη της απόδοσης των μοντέλων αναφορικά με την απόρριψη δεδομένων ταξινόμησης

Σε αυτή την ενότητα θα εξετάσουμε πως μεταβάλλεται η απόδοση των ταξινομητών που δημιουργήσαμε στις προηγούμενες ενότητες (N-gram probabilistic model, vector space model (VSM), latent semantic analysis (LSA)), σε σχέση με την απόρριψη των δεδομένων ταξινόμησης για τα οποία δεν είμαστε πολύ σίγουροι για την κατηγορία στην οποία ανήκουν. Συγκεκριμένα στην περίπτωση του N-gram probabilistic model εξετάζουμε τις πιθανότητες που έχει κάθε υπότιτλος, από το testing set, αναφορικά με κάθε κατηγορία (χιουμοριστική/μη χιουμοριστική). Όμοια στην περίπτωση του VSM και του LSA εξετάζουμε τα similarity scores που έχει κάθε υπότιτλος, από το testing set, αναφορικά με κάθε κατηγορία.

Απορρίπτουμε τους υπότιτλους των οποίων οι λογάριθμοι της πιθανότητας (τύποι 7.4, 7.5, 7.6), που έχουν με κάθε κατηγορία, διαφέρουν λιγότερο από κάποιο κατώφλι. Στην περίπτωση του VSM και του LSA, απορρίπτονται οι υπότιτλοι των οποίων τα similarity scores, μεταξύ των κατηγοριών, έχουν μικρότερη διαφορά από το κατώφλι. Οι υπόλοιποι υπότιτλοι ταξινομούνται κανονικά και με βάση αυτούς γίνεται η εκτίμηση της απόδοσης του ταξινομητή. Βάζοντας όλο και ισχυρότερα κατώφλια στις τιμές των παραπάνω διαφορών και εξετάζοντας σε κάθε βήμα την ορθότητα (accuracy) του ταξινομητή, παίρνουμε μία καμπύλη η οποία μας δείχνει πως μεταβάλλεται το accuracy αναφορικά με την απόρριψη των δεδομένων ταξινόμησης. Τα αποτελέσματα φαίνονται στα σχήματα 7.3, 7.4, 7.5. Σημειώνουμε ότι οι γραφικές αφορούν τα μοντέλα τα οποία έχουν εκπαιδευτεί με το 70% των δεδομένων (ενότητα 7.1 - περίπτωση 1).

7.6 Επιλογή χαρακτηριστικών με βάση την αμοιβαία πληροφορία

Για να δούμε πόσο δυνατή είναι η σχέση κάθε όρου με τις κατηγορίες (χιουμοριστική/μη χιουμοριστική), χρησιμοποιούμε το κριτήριο της αμοιβαίας πληροφορίας. Η αμοιβαία πληροφορία ενός όρου t, με μια κατηγορία c ορίζεται ως εξής:

$$I(t;c) = log(P(t|c)/P(t))$$

Με maximum likelihood estimation, οι παραπάνω πιθανότητες προσεγγίζονται ως εξής:

$$P(t|c) = \frac{\textbf{Αριθμός των εμφανίσεων του όρου t στο κείμενο της κατηγορίας c}}{\textbf{Αριθμός των εμφανίσεων όλων των όρων στο κείμενο της κατηγορίας c}}$$

$$P(t) = \frac{\textbf{Αριθμός των εμφανίσεων του όρου t στην συλλογή των κειμένων}}{\textbf{Αριθμός των εμφανίσεων όλων των όρων στην συλλογή των κειμένων}}$$

Όσο μεγαλύτερη είναι η αμοιβαία πληροφορία του όρου με την κατηγορία, τόσο στενότερα συνδεδεμένος είναι όρος με την κατηγορία. Εάν ένας όρος t δεν εμφανίζεται σε μία κατηγορία c, τότε η αμοιβαία πληροφορία του όρου με την κατηγορία είναι $-\infty$, αφού $P(t|c) = 0$. Πρακτικά στον όρο τίθεται ένα πολύ μικρό βάρος. Επίσης αφαιρούνται οι όροι οι οποίοι οι οποίοι εμφανίζονται το πολύ μια φορά σε κάθε κατηγορία, αφού δεν συμβάλουν ουσιαστικά στην διακριτοποίηση των κατηγοριών μας. Συγκρίνοντας τα scores αμοιβαίας πληροφορίας κάθε όρου με κάθε κατηγορία, επιλέγουμε σαν "καλούς", τους όρους που έχουν μεγάλη διαφορά στα scores αμοιβαίας πληροφορίας μεταξύ των κατηγοριών. Βάζοντας όλο και μεγαλύτερα κατώφλια (thresholds) στην διαφορά, επιλέγουμε όλο και λιγότερους (όμως "δυνατότερους") όρους. Τέλος σε κάθε βήμα επανεκπαιδεύουμε τα μοντέλα μας με τα καινούργια features, και ταξινομούμε τα δεδομένα μας. Η μελέτη της ορθότητας (accuracy) των ταξινομητών αναφορικά με τον αριθμό των training features παρουσιάζεται στα σχήματα 7.6, 7.8, 7.10.

Όσο λιγότερα features χρησιμοποιούμε για την εκπαίδευση των μοντέλων μας, τόσο περισσότεροι υπότιτλοι από τα testing δεδομένα, θα έχουν μηδενική ομοιότητα

και με τα δύο κείμενα της συλλογής μας. Ο λόγος είναι ότι τα features που εξάγονται από τους υπότιτλους προς ταξινόμηση, είναι πολύ πιθανό να μην υπάρχουν στον term document matrix των μοντέλων μας. Οι υπότιτλοι που έχουν μηδενική ομοιότητα και με τα δύο κείμενα της συλλογής μας απορρίπτονται. Η μελέτη της απόρριψης, αναφορικά με τον αριθμό των training features, παρουσιάζεται στα σχήματα 7.7, 7.9, 7.11. Να σημειώσουμε ότι η παραπάνω διαδικασία υλοποιήθηκε με βάση τα training/testing sets του πίνακα 7.2.

7.7 Αποτελέσματα

Στην ενότητα αυτή, συγκεντρώνουμε τα αποτελέσματα της ταξινόμησης των υποτίτλων, σύμφωνα με τα μοντέλα που περιγράφηκαν στις προηγούμενες ενότητες του κεφαλαίου. Θυμίζουμε ότι δημιουργήθηκαν τρία είδη ταξινομητών. Ό πρώτος δημιουργήθηκε σύμφωνα με το N-gram πιθανοτικό μοντέλο, ο δεύτερος σύμφωνα με το μοντέλο διανυσματικού χώρου (Vector Space Model - VSM), και ο τρίτος σύμφωνα με την λανθάνουσα σημασιολογική ανάλυση (Latent Semantic Analysis - LSA). Επίσης παρουσιάζονται τα αποτελέσματα της μελέτης της ορθότητας των ταξινομητών, αναφορικά με την απόρριψη των δεδομένων ταξινόμησης για τα οποία υπάρχει υψηλή αβεβαιότητα για την κατηγορία στην οποία ανήκουν. Τέλος παρουσιάζονται τα αποτελέσματα της μελέτης της ορθότητας των ταξινομητών, όταν γίνεται επιλογή χαρακτηριστικών, αναφορικά με το πλήθος των χαρακτηριστικών που χρησιμοποιούνται στην εκπαίδευση των μοντέλων (training features).

Έστω tp ο αριθμός των χιουμοριστικών υποτίτλων που ταξινομήθηκαν σωστά, και fn ο αριθμός των χιουμοριστικών υποτίτλων που ταξινομήθηκαν λάθος. Αντίστοιχα συμβολίζουμε με tn (fp) τον αριθμό των μη χιουμοριστικών υποτίτλων που ταξινομήθηκαν σωστά (λάθος). Τα αποτελέσματα της ταξινόμησης κάθε πειράματος παρουσιάζονται με πίνακες της παρακάτω μορφής:

true classes/hypothesized classes	humorous	non humorous
humorous	tp	fn
non humorous	fp	tn

Για την εκτίμηση των αποτελεσμάτων μας θα χρησιμοποιήσουμε τρείς μετρικές

(Ορθότητα (Accuracy), ακρίβεια (Precision), ανάκληση (recall)):

1. $Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$
2. $Precision = \frac{tp}{tp + fp}$
3. $Recall = \frac{tp}{tp + fn}$

Η ορθότητα του ταξινομητή υπολογίζεται ως το κλάσμα του αριθμού των σωστών αποφάσεων προς το συνολικό πλήθος των αποφάσεων, δηλαδή μας φανερώνει το ποσοστό των αποφάσεων οι οποίες είναι σωστές. Όμως για την εκτίμηση των αποτελεσμάτων δεν αρκεί μόνο αυτό. Πρέπει να μελετηθεί η συμπεριφορά του ταξινομητή ως προς την μειοψηφική κατηγορία (δηλαδή τους χιουμοριστικούς υπότιτλους που αποτελούν το 45.9% των δεδομένων μας (πίνακας 7.1)). Ουσιαστικά θέλουμε να δούμε αν ο ταξινομητής τείνει να ταξινομεί τα δεδομένα στην πολυπληθέστερη κατηγορία (δηλαδή αν είναι προκατειλημμένος (bias)). Για τον σκοπό αυτό υπολογίζεται η ακρίβεια και η ανάκληση του ταξινομητή ως προς την κατηγορία των χιουμοριστικών υποτίτλων.

Τέλος υπολογίζεται, για κάθε πείραμα, το ποσοστό απόρριψης των ταξινομητών. Απόρριψη ενός υπότιτλου γίνεται, στην περίπτωση του VSM και του LSA, όταν δεν υπάρχει κανένας όρος του υπότιτλου στα μοντέλα που έχουμε εκπαιδεύσει. Στην περίπτωση αυτή δεν υπάρχει κανένας όρος του υπότιτλου στον πίνακα όρων-κειμένων των μοντέλων, με αποτέλεσμα μηδενική ομοιότητα και με τις δύο κατηγορίες. Επίσης ένας υπότιτλος απορρίπτεται στην σπάνια περίπτωση που παρουσιάσει ίδια ομοιότητα, μη μηδενική, και με τις δύο κατηγορίες. Στο N-gram πιθανοτικό μοντέλο εμπλέκονται οι a-priori πιθανότητες των κατηγοριών. Οπότε αν ένας υπότιτλος εμφανίσει το ίδιο score, και με τις δύο κατηγορίες, ταξινομείται στην πολυπληθέστερη κατηγορία. Έτσι στο N-gram πιθανοτικό μοντέλο έχουμε μηδενική απόρριψη δεδομένων. Όμως και στα άλλα μοντέλα (VSM, LSA), τα ποσοστά απόρριψης είναι μηδαμινά (της τάξης του 0.8%).

Ως **baseline** για τις μετρήσεις μας θεωρούμε το ποσοστό της πολυπληθέστερης κατηγορίας. Από τον πίνακα 7.1 βλέπουμε ότι η πολυπληθέστερη κατηγορία είναι οι μη χιουμοριστικοί υπότιτλοι, οι οποίοι είναι το **54.1**% των δεδομένων.

7.7.1 Πειράματα "70% - 30%" και Cross Validation

Στους πίνακες 7.8, 7.9, 7.10, παρουσιάζονται τα αποτελέσματα της ταξινόμησης, για το N-gram πιθανοτικό μοντέλο, το μοντέλο διανυσματικού χώρου, και της λανθάνουσας σημασιολογικής ανάλυσης αντίστοιχα, στην περίπτωση του "70%-30%" πειράματος (ενότητα 7.1 - περίπτωση 1). Τα αποτελέσματα της ορθότητας (accuracy) κάθε μοντέλου βρίσκονται πάνω από το baseline. Δεν παρατηρείται παρατηρείται ιδιαίτερη βελτίωση της ορθότητας αναφορικά με την αύξηση της τάξης των μοντέλων. Επίσης η ορθότητα κάθε ταξινομητή, ασχέτως με το μοντέλο που χρησιμοποιείται (N-gram, VSM, LSA), κυμαίνεται στα ίδια επίπεδα (56.5% με 57%).

<table>
<tr><td>Model</td><td>misclassification tables</td></tr>
<tr><td>unigram</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>554 (tp)</td><td>278 (fn)</td></tr>
<tr><td>non humorous</td><td>485 (fp)</td><td>440 (tn)</td></tr>
</table>
Precision = 0.53321
Recall = 0.66587
Accuracy = 0.56574
#Rejected subtitles = 0 (Rejection = 0%)
</td></tr>
<tr><td>bigram</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>443 (tp)</td><td>389 (fn)</td></tr>
<tr><td>non humorous</td><td>367 (fp)</td><td>558 (tn)</td></tr>
</table>
Precision = 0.54691
Recall = 0.53245
Accuracy = 0.56972
#Rejected subtitles = 0 (Rejection = 0%)
</td></tr>
<tr><td>trigram</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>446 (tp)</td><td>386 (fn)</td></tr>
<tr><td>non humorous</td><td>378 (fp)</td><td>547 (tn)</td></tr>
</table>
Precision = 0.54126
Recall = 0.53606
Accuracy = 0.56517
#Rejected subtitles = 0 (Rejection = 0%)
</td></tr>
</table>

Πίνακας 7.8: Classification results of "70%-30%" experiment (method: N-gram probabilistic model)

Παρατηρείται, από τον πίνακα 7.8, ότι τιμή της ανάκλησης (recall) του ταξινομητή είναι αρκετά υψηλή στην περίπτωση του unigram μοντέλου (66.6%), ενώ στην περίπτωση του bigram και trigram είναι αρκετά χαμηλότερη (53.2% και 53.6% αντίστοιχα). Η ακρίβεια (precision) δεν αλλάζει ιδιαίτερα με την τάξη του μοντέλου και κυμαίνεται στο 54% (53.3% για το unigram, 54.7% για το bigram, και 54.1% για το trigram).

<table>
<tr><td>Features</td><td>misclassification tables</td></tr>
<tr><td>unigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>485 (tp)</td><td>336 (fn)</td></tr>
<tr><td>non humorous</td><td>421 (fp)</td><td>504 (tn)</td></tr>
</table>
Precision = 0.53532
Recall = 0.59074
Accuracy = 0.56644
#Rejected subtitles = 11 (Rejection = 0.6%)
</td></tr>
<tr><td>unigrams/bigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>479 (tp)</td><td>342 (fn)</td></tr>
<tr><td>non humorous</td><td>411 (fp)</td><td>514 (tn)</td></tr>
</table>
Precision = 0.5382
Recall = 0.58343
Accuracy = 0.56873
#Rejected subtitles = 11 (Rejection = 0.6%)
</td></tr>
<tr><td>unigrams/bigrams/trigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>468 (tp)</td><td>353 (fn)</td></tr>
<tr><td>non humorous</td><td>396 (fp)</td><td>529 (tn)</td></tr>
</table>
Precision = 0.54167
Recall = 0.57004
Accuracy = 0.57102
#Rejected subtitles = 11 (Rejection = 0.6%)
</td></tr>
</table>

Πίνακας 7.9: Classification results of "70%-30%" experiment (method: VSM)

Σε αντίθεση με το N-gram πιθανοτικό μοντέλο, στο μοντέλο διανυσματικού χώρου η τιμή της ανάκλησης του ταξινομητή δεν αλλάζει ιδιαίτερα (παρότι παρατηρείται και πάλι μικρή μείωση της ανάκλησης) με την τάξη του μοντέλου (59%

για unigram όρους, 58.3% για το unigram/bigram όρους, και 57% για το unigram/bigram/trigram όρους). Όσον αφορά την ακρίβεια και πάλι κυμαίνεται στο 54%.

<table>
<tr><td>Features</td><td>misclassification tables</td></tr>
<tr><td>unigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>484 (tp)</td><td>337 (fn)</td></tr>
<tr><td>non humorous</td><td>419 (fp)</td><td>506 (tn)</td></tr>
</table>
Precision = 0.53599
Recall = 0.58952
Accuracy = 0.56701
#Rejected subtitles = 11 (Rejection = 0.6%)
</td></tr>
<tr><td>unigrams/bigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>476 (tp)</td><td>345 (fn)</td></tr>
<tr><td>non humorous</td><td>407 (fp)</td><td>518 (tn)</td></tr>
</table>
Precision = 0.53907
Recall = 0.57978
Accuracy = 0.5693
#Rejected subtitles = 11 (Rejection = 0.6%)
</td></tr>
<tr><td>unigrams/bigrams/trigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>468 (tp)</td><td>353 (fn)</td></tr>
<tr><td>non humorous</td><td>396 (fp)</td><td>529 (tn)</td></tr>
</table>
Precision = 0.54167
Recall = 0.57004
Accuracy = 0.57102
#Rejected subtitles = 11 (Rejection = 0.6%)
</td></tr>
</table>

Πίνακας 7.10: Classification results of "70%-30%" experiment (method: LSA)

Όπως και στο μοντέλο διανυσματικού χώρου, έτσι και στην λανθάνουσα σημασιολογική ανάλυση, η ανάκληση του ταξινομητή δεν αλλάζει ιδιαίτερα με την τάξη του μοντέλου (59% για unigram όρους, 58% για το unigram/bigram όρους, και 57% για το unigram/bigram/trigram όρους). Όσον αφορά την ακρίβεια και πάλι κυμαίνεται στο 54%.

Στους πίνακες 7.11, 7.12, 7.13, παρουσιάζονται τα αποτελέσματα της ταξινόμησης, για το N-gram πιθανοτικό μοντέλο, το μοντέλο διανυσματικού χώρου, και της

λανθάνουσας σημασιολογικής ανάλυσης αντίστοιχα, στην περίπτωση του cross validation πειράματος (ενότητα 7.1 - περίπτωση 2). Τα αποτελέσματα της ορθότητας (accuracy) κάθε μοντέλου βρίσκονται, όπως και προηγουμένως, πάνω από το baseline. Όπως και στα αποτελέσματα του προηγούμενου πειράματος, έτσι και στο cross validation πείραμα, δεν παρατηρείται ιδιαίτερη βελτίωση της ορθότητας αναφορικά με την αύξηση της τάξης των μοντέλων. Στο cross validation πείραμα, το N-gram πιθανοτικό μοντέλο παρουσιάζει λίγο μεγαλύτερες τιμές στην ορθότητα (που κυμαίνονται στο 56.5%), σε σχέση με τις τιμές στο μοντέλο διανυσματικού χώρου και την λανθάνουσα σημασιολογική ανάλυση (που κυμαίνονται στο 55.5%).

<table>
<tr><td>Model</td><td>misclassification tables</td></tr>
<tr><td>unigram</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1706 (tp)</td><td>1114 (fn)</td></tr>
<tr><td>non humorous</td><td>1551 (fp)</td><td>1774 (tn)</td></tr>
</table>
Precision = 0.52379
Recall = 0.60496
Accuracy = 0.56631
#Rejected subtitles = 0 (Rejection = 0%)
</td></tr>
<tr><td>bigram</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1389 (tp)</td><td>1431 (fn)</td></tr>
<tr><td>non humorous</td><td>1241 (fp)</td><td>2084 (tn)</td></tr>
</table>
Precision = 0.52814
Recall = 0.49255
Accuracy = 0.56517
#Rejected subtitles = 0 (Rejection = 0%)
</td></tr>
<tr><td>trigram</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1399 (tp)</td><td>1421 (fn)</td></tr>
<tr><td>non humorous</td><td>1271 (fp)</td><td>2054 (tn)</td></tr>
</table>
Precision = 0.52397
Recall = 0.4961
Accuracy = 0.56192
#Rejected subtitles = 0 (Rejection = 0%)
</td></tr>
</table>

Πίνακας 7.11: Overall classification results of cross-validation experiment (method: N-gram probabilistic model)

Όπως φαίνεται από τον παραπάνω πίνακα, παρατηρείται και πάλι (στο N-gram

πιθανοτικό μοντέλο) δραστική μείωση της ανάκλησης όταν περνάμε από το unigram στο bigram και στο trigram μοντέλο. Η τιμή της ακρίβειας δεν μεταβάλλεται ιδιαίτερα, αναφορικά με την τάξη του μοντέλου.

<table>
<tr><td>Features</td><td>misclassification tables</td></tr>
<tr><td>unigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1572 (tp)</td><td>1207 (fn)</td></tr>
<tr><td>non humorous</td><td>1533 (fp)</td><td>1781 (tn)</td></tr>
</table>
Precision = 0.50628
Recall = 0.56567
Accuracy = 0.5503
#Rejected subtitles = 52 (Rejection = 0.8%)
</td></tr>
<tr><td>unigrams/bigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1571 (tp)</td><td>1208 (fn)</td></tr>
<tr><td>non humorous</td><td>1507 (fp)</td><td>1807 (tn)</td></tr>
</table>
Precision = 0.5104
Recall = 0.56531
Accuracy = 0.55441
#Rejected subtitles = 52 (Rejection = 0.8%)
</td></tr>
<tr><td>unigrams/bigrams/trigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1548 (tp)</td><td>1231 (fn)</td></tr>
<tr><td>non humorous</td><td>1481 (fp)</td><td>1833 (tn)</td></tr>
</table>
Precision = 0.51106
Recall = 0.55703
Accuracy = 0.5549
#Rejected subtitles = 52 (Rejection = 0.8%)
</td></tr>
</table>

Πίνακας 7.12: Overall classification results of cross-validation experiment (method: VSM)

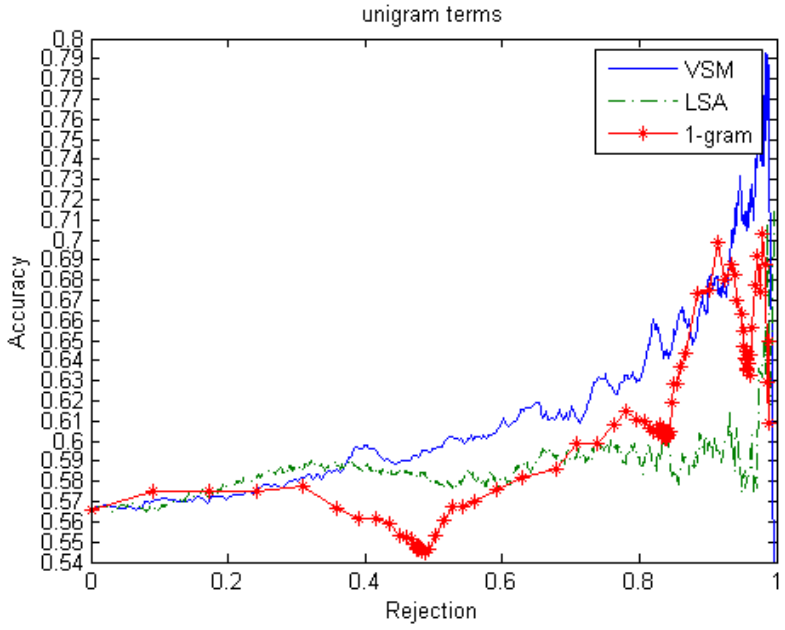
Όπως παρατηρούμε από τον παραπάνω πίνακα τόσο οι τιμές της ανάκλησης, όσο και οι τιμές της ακρίβειας, στο μοντέλο του διανυσματικού χώρου, δεν μεταβάλλονται κατά σημαντικά ποσοστά, αναφορικά με την τάξη του μοντέλου. Η ίδια παρατήρηση προκύπτει και από τον επόμενο πίνακα που αφορά την λανθάνουσα σημασιολογική ανάλυση.

<table>
<tr><td>Features</td><td>misclassification tables</td></tr>
<tr><td>unigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1572 (tp)</td><td>1207 (fn)</td></tr>
<tr><td>non humorous</td><td>1526 (fp)</td><td>1788 (tn)</td></tr>
</table>
Precision = 0.50742
Recall = 0.56567
Accuracy = 0.55145
#Rejected subtitles = 52 (Rejection = 0.8%)
</td></tr>
<tr><td>unigrams/bigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1560 (tp)</td><td>1219 (fn)</td></tr>
<tr><td>non humorous</td><td>1495 (fp)</td><td>1819 (tn)</td></tr>
</table>
Precision = 0.51064
Recall = 0.56135
Accuracy = 0.55457
#Rejected subtitles = 52 (Rejection = 0.8%)
</td></tr>
<tr><td>unigrams/bigrams/trigrams</td><td>
<table>
<tr><td>true classes/hypothesized classes</td><td>humorous</td><td>non humorous</td></tr>
<tr><td>humorous</td><td>1533 (tp)</td><td>1246 (fn)</td></tr>
<tr><td>non humorous</td><td>1453 (fp)</td><td>1861 (tn)</td></tr>
</table>
Precision = 0.5134
Recall = 0.55164
Accuracy = 0.55703
#Rejected subtitles = 52 (Rejection = 0.8%)
</td></tr>
</table>

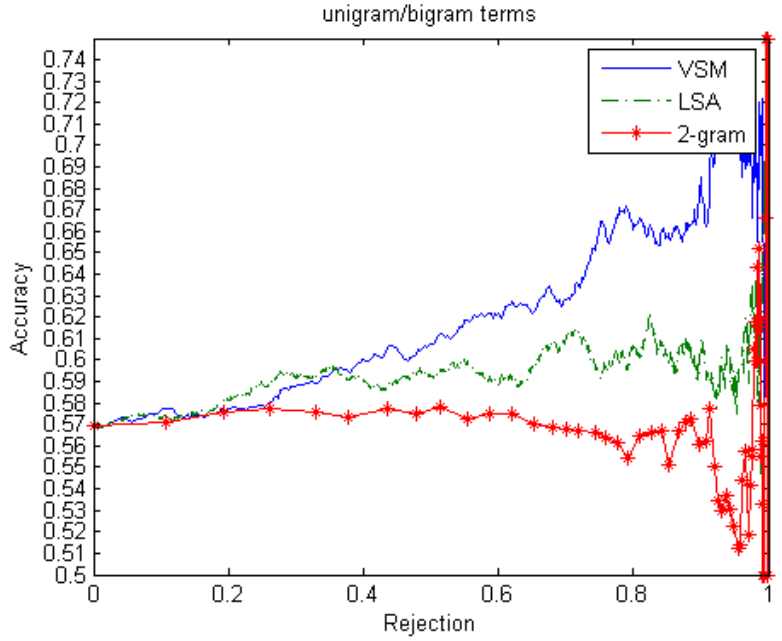
Πίνακας 7.13: Overall classification results of cross-validation experiment (method: LSA)

7.7.2 Μελέτη της ορθότητας των ταξινομητών, αναφορικά με την απόρριψη δεδομένων ταξινόμησης

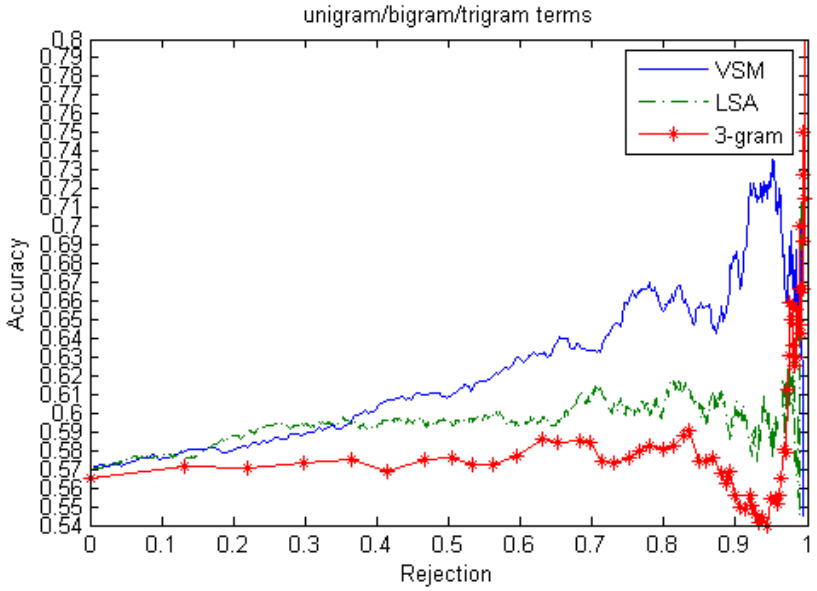
Στα επόμενα σχήματα (7.3, 7.4, 7.5), παρουσιάζεται η μεταβολή της ορθότητας (accuracy) των ταξινομητών, αναφορικά με την απόρριψη των υποτίτλων για τους οποίους υπάρχει υψηλή αβεβαιότητα για την κατηγορία στην οποία ανήκουν (ενότητα 7.5). Παρατηρείται και και στα τρία σχήματα, αύξηση της ορθότητας των ταξινομητών με την αύξηση της απόρριψης των υποτίτλων. Το μοντέλο του διανυσματικού χώρου παρουσιάζει καλύτερη συμπεριφορά, και στα τρία σχήματα, σε σχέση με το N-gram πιθανοτικό μοντέλο και την λανθάνουσα σημασιολογική ανάλυση.



Σχήμα 7.3: Μεταβολή ορθότητας αναφορικά με την απόρριψη δεδομένων ταξινόμησης (Τα μοντέλα χρησιμοποιούν unigram όρους)



Σχήμα 7.4: Μεταβολή ορθότητας αναφορικά με την απόρριψη δεδομένων ταξινόμησης (Τα μοντέλα χρησιμοποιούν unigram/bigram όρους)

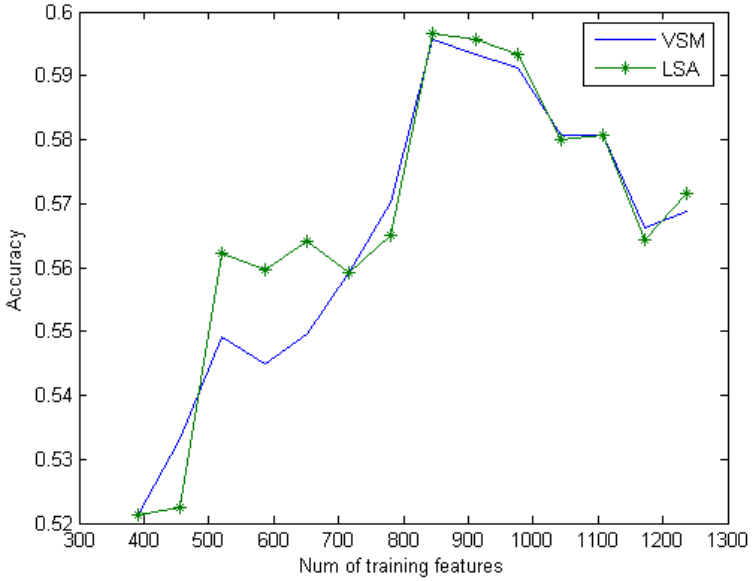


Σχήμα 7.5: Μεταβολή ορθότητας αναφορικά με την απόρριψη δεδομένων ταξινόμησης (Τα μοντέλα χρησιμοποιούν unigram/bigram/trigram όρους)

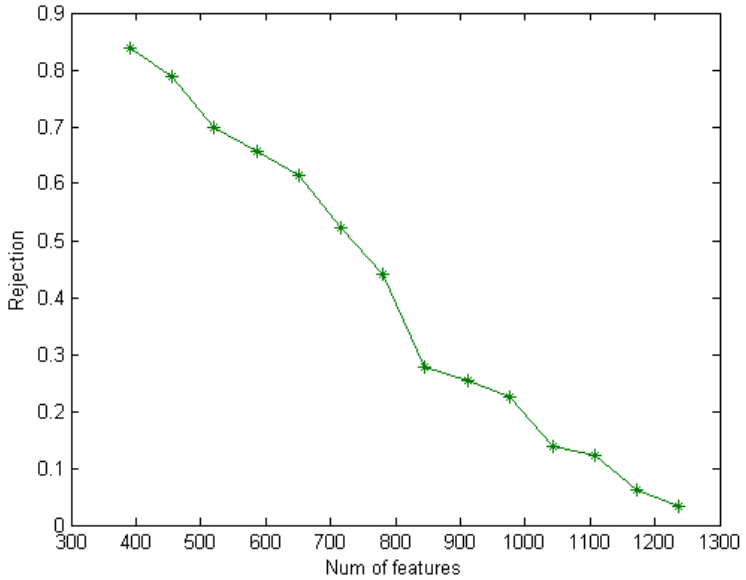
7.7.3 Μελέτη της ορθότητας των ταξινομητών, αναφορικά με το πλήθος των χαρακτηριστικών εκπαίδευσης

Στα σχήματα 7.6, 7.8, 7.10 παρουσιάζεται η μεταβολή της ορθότητας των ταξινομητών, αναφορικά με το πλήθος των χαρακτηριστικών που χρησιμοποιούνται στην εκπαίδευση τους. Η επιλογή των χαρακτηριστικών γίνεται με κριτήριο την αμοιβαία πληροφορία (ενότητα 7.6). Παρατηρείται και στα τρία σχήματα μία πτώση της ορθότητας των ταξινομητών στα άκρα, και μία μεγάλη αύξηση στην ενδιάμεση περιοχή. Αυτό συμβαίνει για τον εξής λόγο: Όταν ο αριθμός των χαρακτηριστικών εκπαίδευσης (training features) είναι πολύ μικρός, τα μοντέλα δεν εκπαιδεύονται με την απαιτούμενη πληροφορία με αποτέλεσμα την αύξηση των λαθών ταξινόμησης. Από την άλλη πλευρά, όταν αριθμός των training features είναι μεγάλος, δεν έχουν απορριφθεί ακόμη τα αδύναμα χαρακτηριστικά με αποτέλεσμα μειωμένη απόδοση. Στο ενδιάμεσο, όπου υπάρχουν αρκετά και δυνατά χαρακτηριστικά, τα αποτελέσματα είναι ικανοποιητικά, με την ορθότητα να φτάνει στο **60**% (**5.9**% πάνω από το **baseline**).

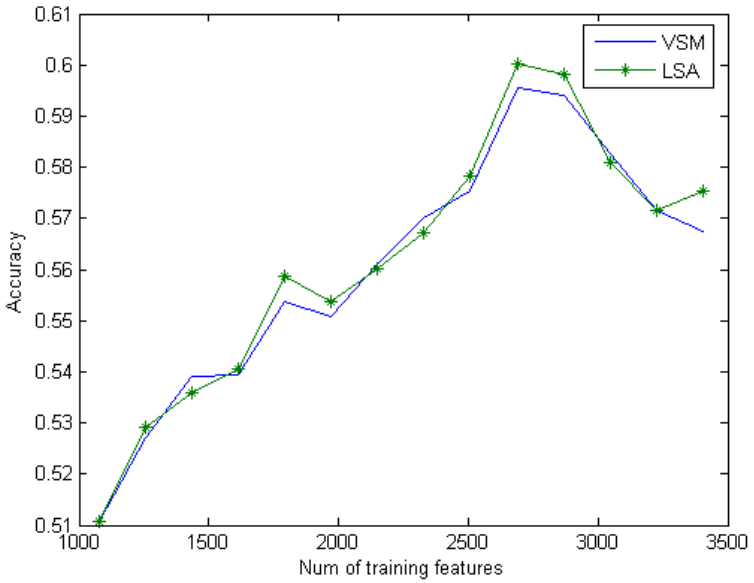
Στα σχήματα 7.7, 7.9, 7.11 φαίνεται πως μεταβάλλεται η απόρριψη των δεδομένων ταξινόμησης, αναφορικά με το πλήθος των χαρακτηριστικών που χρησιμοποιούνται στην εκπαίδευση των μοντέλων. Όπως φαίνεται και από τα σχήματα, όσο λιγότερα χαρακτηριστικά χρησιμοποιούμε στην εκπαίδευση των μοντέλων, τόσο μεγαλύτερη πιθανότητα έχει κάποιος υπότιτλος να παρουσιάζει μηδενική ομοιότητα και με τις δύο κατηγορίες (αφού οι όροι του υπότιτλου δεν θα υπάρχουν στον πίνακα όρων-κειμένων), με αποτέλεσμα την αυξημένη απόρριψη. Πρέπει να σημειωθεί ότι οι καμπύλες της απόρριψης των δεδομένων συμπίπτουν για τα VSM και LSA, αφού τα μοντέλα αυτά έχουν τον ίδιο πίνακα όρων - κειμένων. Έτσι παρουσιάζεται μία καμπύλη σε κάθε σχήμα η οποία είναι κοινή για το VSM και το LSA.



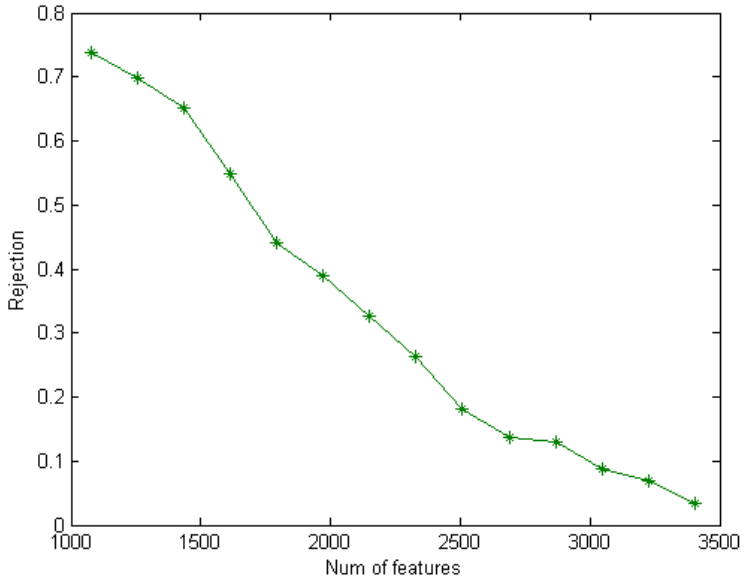
Σχήμα 7.6: Μεταβολή ορθότητας αναφορικά με πλήθος των χαρακτηριστικών εκπαίδευσης (Τα μοντέλα χρησιμοποιούν unigram όρους)



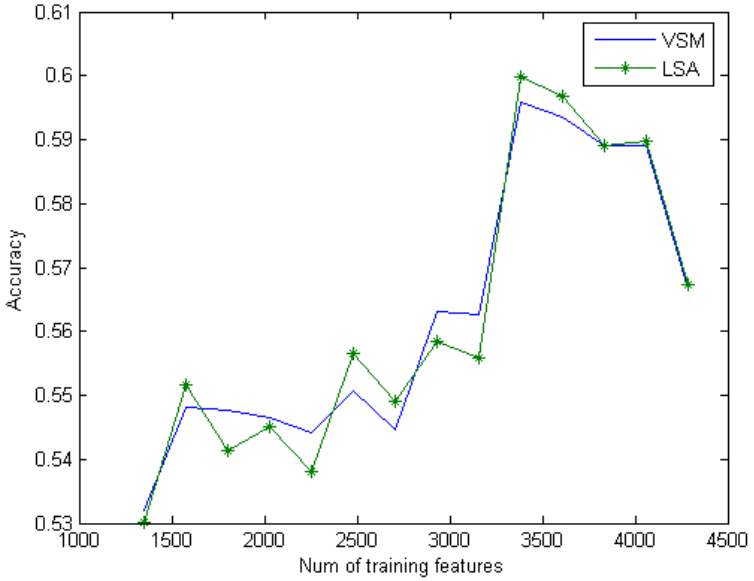
Σχήμα 7.7: Μεταβολή απόρριψης δεδομένων αναφορικά με πλήθος των χαρακτηριστικών εκπαίδευσης (Τα μοντέλα χρησιμοποιούν unigram όρους)



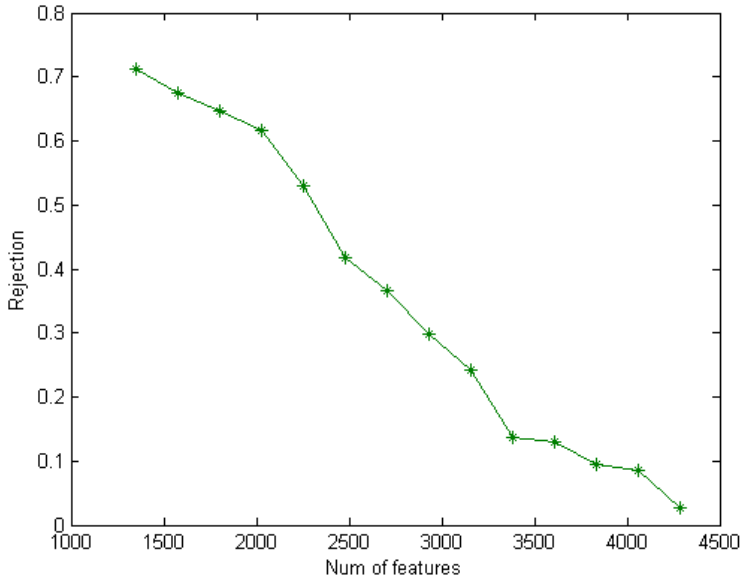
Σχήμα 7.8: Μεταβολή ορθότητας αναφορικά με πλήθος των χαρακτηριστικών εκπαίδευσης (Τα μοντέλα χρησιμοποιούν unigram/bigram όρους)



Σχήμα 7.9: Μεταβολή απόρριψης δεδομένων αναφορικά με πλήθος των χαρακτηριστικών εκπαίδευσης (Τα μοντέλα χρησιμοποιούν unigram/bigram όρους)



Σχήμα 7.10: Μεταβολή ορθότητας αναφορικά με πλήθος των χαρακτηριστικών εκπαίδευσης (Τα μοντέλα χρησιμοποιούν unigram/bigram/trigram όρους)



Σχήμα 7.11: Μεταβολή απόρριψης δεδομένων αναφορικά με πλήθος των χαρακτηριστικών εκπαίδευσης (Τα μοντέλα χρησιμοποιούν unigram/bigram/trigram όρους)

Κεφάλαιο 8

Σύνοψη

8.1 Συμπεράσματα

Η μοντελοποίηση του χιούμορ είναι ένα αρκετά πολύπλοκο και δύσκολο θέμα. Η χιουμοριστική γλώσσα χρησιμοποιεί πολύπλοκες, ασυνεπείς, και διφορούμενες συντακτικές και σημασιολογικές δομές, οι οποίες απαιτούν βαθιά σημασιολογική εξήγηση-ερμηνεία.

Στην δική μας εργασία χρησιμοποιήθηκαν μόνο γλωσσικά χαρακτηριστικά, οπότε σαν μέτρο σύγκρισης, με τον ταξινομητή των A. Purandare και D. Litman [2], πρέπει να λάβουμε την ακρίβεια 61.14% που αφορά τα γλωσσικά χαρακτηριστικά (πίνακας 5.8). Βέβαια να σημειωθεί ότι στα χαρακτηριστικά που δίνουν ακρίβεια 61.14%, επιπλέον των γλωσσικών, συμπεριλαμβάνεται και το χαρακτηριστικό του μήκους της πρότασης του ηθοποιού (Turn Length), που προσδίδει επιπλέον πληροφορία.

Όσον αφορά την δική μας εργασία πρέπει να ληφθούν υπόψιν τα εξής θέματα: Πρώτον, η εξαγωγή των χιουμοριστικών υποτίτλων γίνεται με αυτόματο τρόπο (κεφάλαιο 6) εισάγοντας ένα περιθώριο λάθους της τάξεως του 4.5% (πίνακας 6.9). Από την άλλη πλευρά, στην εργασία των A. Purandare και D. Litman, η επισήμανση των χιουμοριστικών speaker turns γίνεται χειροκίνητα με μηδενικό περιθώριο λάθους. Με άλλα λόγια στα γλωσσικά δεδομένα, που χρησιμοποιούνται στο κεφάλαιο 7 για την αυτόματη αναγνώριση του χιούμορ, έχει προστεθεί θόρυβος (υπότιτλοι που θεωρούνται χιουμοριστικοί ενώ δεν είναι) της τάξης του 4.5%. Δεύτερον, στην εργασία

μας ταξινομούμε υπότιτλους (σε κατηγορίες χιουμοριστικών/μη χιουμοριστικών). Σε έναν υπότιτλο μερικές φορές συμμετέχουν παραπάνω από ένας ηθοποιοί. Αντίθετα, στην εργασία των A. Purandare και D. Litman, ταξινομούνται speaker turns. Με τον όρο speaker turn (μιλώντας για γλωσσικά χαρακτηριστικά-κείμενο) εννοούμε το κείμενο του ηθοποιού στην σειρά του. Δηλαδή στο speaker turn μιλάει ένας ηθοποιός. Συνήθως το χιούμορ, στις χιουμοριστικές σειρές (στην περίπτωση μας στην σειρά Friends), εκφράζεται από έναν ηθοποιό την φορά. Παραδείγματος χάριν, στον υπότιτλο :
-Not every morning
-Making it worse!
, συμμετέχουν δύο ηθοποιοί.

Εφόσον ο υπότιτλος είναι χιουμοριστικός, όλα τα γλωσσικά χαρακτηριστικά του (not, every, morning, making, it, worse) θα περάσουν στην κατηγορία του χιούμορ. Όμως στην συγκεκριμένη περίπτωση το γέλιο προκαλείται από την δεύτερη πρόταση, και θα έπρεπε να περάσουν στην κατηγορία του χιούμορ μόνο τα χαρακτηριστικά της δεύτερης πρότασης (making, it, worse). Όπως φαίνεται από το παραπάνω παράδειγμα, τέτοιες περιπτώσεις εισάγουν επιπλέον θόρυβο, κάνοντας το εγχείρημα της αναγνώρισης ακόμα πιο δύσκολο.

Παρά τις δύο αδυναμίες που περιγράφηκαν παραπάνω, τα αποτελέσματα μας (ενότητα 7.7) είναι αξιόλογα. Σε όλα τα πειράματα η ακρίβεια των ταξινομητών μας βρίσκεται πάνω από το baseline. Ειδικότερα, στην περίπτωση που γίνεται επιλογή χαρακτηριστικών (διαγράμματα 7.6 έως 7.11) η ακρίβεια φτάνει στο 60%, ποσοστό πολύ κοντινό με το 61.14% που αναφέραμε σαν μέτρο σύγκρισης.

8.2 Μελλοντικός σχεδιασμός

Στον μελλοντικό σχεδιασμό για περαιτέρω βελτίωση των αποτελεσμάτων αξίζει να σημειώσουμε τα επόμενα θέματα. Πρώτον, την δεικτοδότηση των speaker turns αντί των υποτίτλων. Δεύτερον, θα είχε ενδιαφέρον η πηγή των χιουμοριστικών δεδομένων να είχε ποικιλομορφία. Στην εργασία μας η πηγή των χιουμοριστικών και μη χιουμοριστικών εκφράσεων ήταν μία, κάνοντας τα δεδομένα να έχουν μεγάλη

ομοιογένεια, με αποτέλεσμα τον δύσκολο διαχωρισμό τους. Τέλος θα είχε ενδιαφέρον και η χρησιμοποίηση και άλλων, πέραν των γλωσσικών, χαρακτηριστικών στην πειραματική διαδικασία. Τέτοια χαρακτηριστικά θα μπορούσαν να είναι ακουστικής - προσωδιακής φύσεως ή περισσότερο πολύπλοκα σημασιολογικά και πραγματολογικά χαρακτηριστικά όπως αμφισημία (ambiguity), και ασυμφωνία (incongruity).

Βιβλιογραφία

[1] Μ. Χαλκίδη, Μ.Βαζιργιάννης. *Εξόρυξη γνώσης από βάσεις δεδομένων και τον παγκόσμιο ιστό*. Τυπωθήτω, Αθήνα 2006.

[2] Amruta Purandare, Diane Litman. *Humor: Prosody Analysis and Automatic Recognition for friends*. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 208-215, Sydney, July 2006.

[3] C. Alm, D. Roth, R. Sproat. *Emotions from text: Machine learning for text-based emotion prediction*. In Proceedings of HLT/EMNLP, Vancouver, CA, 2005.

[4] S. Attardo. *Linguistic Theory of Humor*. Mounton de Gruyter, Berlin, 1994.

[5] H. Pain, A. Waller, D. O'Mara. *Computational humor*. IEEE Intelligent Systems, March-April, 2006.

[6] R. Mihalcea, C.Strapparava. *Making computers laugh: Investigations in automatic humor recognition*. In Proceedings of HLT/EMNLP, Vancouver, CA, 2005.

[7] M. Mulder, A. Nijholt. *Humor research: State of the art*. Technical Report 34, CTIT Technical Report Series, 2002.

[8] O. Stock, C.Strapparava. *Hahaacronym: A computational humor system*. In Proceedings of ACL Interactive Poster and Demonstration Session, pages 113-116, Ann Arbor, MI, 2005.

[9] J. Taylor, L. Mazlack. *Computationally recognizing wordplay in jokes*. In Proceedings of the CogSci 2004, Chicago, IL, 2004.

[10] Daniel Jurafsky, James H.Martin. *Speech and Language Processing*. Prentice Hall, 2000.

[11] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[12] R. Duda, D. Stork, P. Hart. *Pattern Classification*. John Wiley & Sons, 2000.

[13] *The CMU-Cambridge Statistical Language Modeling Toolkit v2*. *http* : *//mi.eng.cam.ac.uk/* $\sim$ *prc*14/*toolkit_documentation.html*.

[14] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, Phil Woodland. *The HTK book*. http://htk.eng.cam.ac.uk, 2009.

[15] Dr. Edel Garcia. *Latent Semantic Indexing (LSI) A Fast Track Tutorial*. http://www.miislita.com/information-retrieval-tutorial/latent-semantic-indexing-fast-track-tutorial.pdf, 2006.

[16] *Note Set 4: The EM Algorithm for Gaussian Mixtures*. *http* : *//www.ics.uci.edu/* $\sim$ *smyth/courses/ics*274/*notes*4.*pdf*.

[17] Jeff A. Bilmes. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, April 1998.

[18] Lei Xu, Michael I. Jordan. *On Convergence Properties of the EM Algorithm for Gaussian Mixtures*. ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-1520.pdf, January 1995.